A PERMUTATION-BASED CORRECTION FOR PEARSON'S CHI-SQUARE TEST ON DATA WITH AN IMPUTED COMPLEX OUTCOME / A MODIFIED EM ALGORITHM FOR CONTINGENCY TABLE ANALYSIS WITH MISSING DATA

by

Megan J. Olson Hunt

B.S.T. Mathematics, Education, Winona State University, 2007

B.A. Psychology, Statistics, Winona State University, 2007

Submitted to the Graduate Faculty of
the Department of Biostatistics
Graduate School of Public Health in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Megan J. Olson Hunt

It was defended on

April 3, 2014

and approved by

Gong Tang, PhD, Associate Professor
Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Andriy Bandos, PhD, Assistant Professor
Departments of Biostatistics and Radiology
Graduate School of Public Health, University of Pittsburgh

Maria Mori Brooks, PhD, Associate Professor
Departments of Biostatistics and Epidemiology
Graduate School of Public Health, University of Pittsburgh

Chung-Chou Ho Chang, PhD, Professor
Departments of Medicine and Biostatistics
Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Gong Tang, PhD, Associate Professor
Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

A PERMUTATION-BASED CORRECTION FOR PEARSON'S CHI-SQUARE TEST ON DATA WITH AN IMPUTED COMPLEX OUTCOME / A MODIFIED EM ALGORITHM FOR CONTINGENCY TABLE ANALYSIS WITH MISSING DATA

Megan J. Olson Hunt, PhD

University of Pittsburgh, 2014

ABSTRACT

Studies on human subjects often yield missing data, making progress in this field of inherent public health relevance. Here, two statistical methods are proposed for the analysis of discrete data with missing values. First, when one variable is subject to missingness, it was noted the application of Pearson's chi-square test to singly-imputed data undermines the variability due to imputation, leading to a type-I error rate larger than the nominal level. This research concerns Pearson's test on data with an imputed complex outcome, where one of its components suffers from missing values. Imputation in this context may be performed either directly through conditional imputation of the complex outcome given covariates, or indirectly through conditional imputation of its missing component given the covariates and the other, observed component. Although the latter imputation scheme is shown to be more efficient, an existing adjustment method cannot be extended to this scenario due to the lack of independence amongst the variables constituting the complex outcome. As a result, a novel permutation-based correction method for Pearson's test

is proposed. Simulation studies indicate it provides the nominal rejection rate under the null. Second, a modification of the expectation maximization (EM) algorithm for the analysis of discrete data with missing values is presented. In general, the update in the M-step requires either knowing or modeling the missing-data mechanism. However, misspecification of this mechanism may lead to biased estimates of model parameters. Given consistent initial estimates of the parameters (which may be obtained from an external, complete data set, or by recalling a random sample of subjects), the target function is approximated in the M-step with empirical estimates, allowing for unbiased estimation without specification or modeling of the often intangible missing-data mechanism. Simulation studies show this modified algorithm yields consistent estimates potentially more efficient than the initial estimates, even under non-ignorable missingness.

Keywords: single imputation, discrete data, bias, consistency, efficiency, MNAR, empirical.

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

Thank you to those who contributed to my development as a statistician during my time at the University of Pittsburgh: my dissertation advisor, Dr. Gong Tang; my dissertation committee, Drs. Andriy Bandos, Maria Mori Brooks and Joyce Chang; my mentors and teachers, Drs. Stewart Anderson, Robert Boudreau, Joseph Costantino, Richard Day, Jong-Hyeon Jeong, Gary Marsh, Howard Rockette, Caterina Rosano, Roslyn Stone, Abdus Wahed, Lisa Weissfeld, John Wilson and Ada Youk; and my husband, Sam.

1.0 INTRODUCTION

1.1 MISSING DATA

In the process of data collection, especially on human subjects, missing data may arise for multiple reasons. For example, subjects may move away or have an adverse reaction to (or feel they have received the maximum benefit from) a treatment, such that they no longer wish to participate in a study. When data are collected over time, as in longitudinal studies, this problem can be exacerbated, as participant drop-out tends to increase with an increasing number of follow-up sessions. Consequently, one may see a potentially large decline in sample size for measures taken later in a study. Survey data are also particularly susceptible to missing values, as questions may be deemed too personal or too numerous. In research involving mechanical processes, machines may fail because of experimental conditions, or the failure may be independent of the study parameters.

Missing data may also arise from the study design itself. In some instances, procedures or tests are resource- or monetarily-intensive, or may cause adverse side-effects. As a result, researchers will choose only to collect certain variables for a subset of the subjects, leaving values of those variables missing in the remainder of the sample. Specifically in diagnostic testing, researchers may assume a negative result

on one test implies a negative result on another, such that the latter data point is missing by design. However, given the potential for false negatives, this assumption is not always valid. Often when data are missing − by design or otherwise − imputation is conducted, subsequently allowing the utilization of statistical methods developed for complete, rectangular data sets.

Motivation for imputation stems from the existence of statistical procedures that cannot manage observations with missingness. In these instances, subjects without complete data would be omitted from the analysis entirely. This complete-case analysis results in reduced efficiency, as less information from the data is utilized, and may produce biased estimates if the reason for missingness is not random. For example, if a given treatment has an adverse effect on males more often than females, males may be more likely to drop-out. Subsequently, conclusions no longer apply to a random sample of men, but only to the more robust subset who remained in the study, resulting in questionable external validity. Additionally, this type of drop-out may affect the significance of the effect of gender on the outcome if this loss of data reduced or increased differences across genders. Lastly, missingness may also limit the sample size in certain sub-populations (like genders or ethnicities), such that valid inference cannot be made on these variables.

In general, data may be missing randomly, or the missingness may depend on certain variables, which themselves may be fully-observed or subject to missingness. These concepts are discussed in the following section.

1.2 PATTERNS AND MECHANISMS OF MISSINGNESS

In the study of missing data, *patterns* and *mechanisms* of missingness are important notions that drive theoretical derivations and practical applications. Some issues related to *patterns* of missingness are 1) whether the missingness involves one or more variables, 2) whether or not the data are *monotonically* missing (i.e., if certain variables are measured over time, whether or not missingness at one occasion implies missingness of values thereafter), 3) whether the variable is observable or latent and 4) whether there are certain variables combinations that are never observed together in a final data set, which is often caused by large amounts of missing data (Little and Rubin, 2002). Regarding (2), there exist certain methods applicable to monotonically missing data that are not appropriate for more general patterns, making the former preferable. The last notion is of concern as it causes some parameters not to be estimable.

In contrast, missing data *mechanisms* refer to the underlying cause of the missingness. For example, if data were missing because a subject was ill, moved unexpectedly or got called into work, the missingness does not depend on the variables under study. Such a mechanism is defined by Rubin (1976) as *missing completely at random* (MCAR). Alternatively, missing observations could be related to the actual variables being collected. Of importance, then, is whether the missing data are related to something one did or did not observe (but in theory could have).

Take, for example, a longitudinal study where one collected the age of participants at baseline, then found older people were more likely to drop-out as the study progressed. Since the missing values were related to something that *was* observed (age), the mechanism is defined as *missing at random* (MAR).

Lastly, data may be *missing not at random* (MNAR). Here, the missing data

depend on something one did not observe − the variable with the missingness, and, more specifically, the missing values themselves. As an example, consider measuring income: It is possible people with very high or very low income may not want to report this fact on a survey. Thus, there will be missing income values, and the reason is due to the values themselves (very high/low income). The formal definitions of MCAR, MAR and MNAR as characterized by Rubin (1976) are as follows:

Let $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ represent a data set (matrix), where $\mathbf{Y}_{obs}$ is the observed portion and $\mathbf{Y}_{mis}$ the missing. Define $\mathbf{R}$ as the missing data matrix, such that $r_{ij} = 1$ if y_{ij} is observed for subject i and variable j, 0 otherwise. Then, if the distribution of $\mathbf{R}$ given $\mathbf{Y}$ is denoted $f(\mathbf{R} \,|\, \mathbf{Y}, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ represents the parameters of $\mathbf{R}$, then

$$\begin{aligned}
&\text{MCAR: } f(\mathbf{R} \,|\, \mathbf{Y}, \boldsymbol{\psi}) = f(\mathbf{R} \,|\, \boldsymbol{\psi}) \;\forall\; \mathbf{Y} \\
&\text{MAR: } f(\mathbf{R} \,|\, \mathbf{Y}, \boldsymbol{\psi}) = f(\mathbf{R} \,|\, \mathbf{Y}_{obs}, \boldsymbol{\psi}) \;\forall\; \mathbf{Y} \\
&\text{MNAR: } f(\mathbf{R} \,|\, \mathbf{Y}, \boldsymbol{\psi}) = f(\mathbf{R} \,|\, \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}) \;\forall\; \mathbf{Y}
\end{aligned}$$

The following section provides an overview of missing data techniques, their assumptions with regard to the missingness mechanism, as well as advantages and disadvantages.

1.3 METHODS FOR THE ANALYSIS OF MISSING DATA

As given in Little and Rubin (2002), there are four main methods to deal with missing data: 1) complete-case analysis, 2) weighting, 3) imputation and 4) model-based procedures, which include maximum likelihood (ML), generalized estimating equations and Bayesian methods.

1.3.1 Complete-case analysis

As the name implies, complete-case analysis – the simplest approach to analyzing missing data – uses only those subjects with complete data across all variables. Often the default for certain procedures in statistical packages, the sample size will be reduced to completers only unless otherwise specified.

Although simple, this technique is usually not preferable due to the loss of data, which causes a decrease in precision, affecting inference. Additionally, the potential for bias exists unless the missingness can be verified as MCAR and missing observations are randomly distributed across the sample (Little and Rubin, 2002). As discussed previously, this approach may also limit the sample size within subgroups of a study, resulting in certain parameters that are not estimable. If the amount of missingness is small, the magnitude of these issues may be deemed negligible and the approach acceptable.

1.3.2 Weighting

The goal of weighting is to adjust for potential biases that could otherwise be realized in a complete-case analysis (Little and Rubin, 2002). One example is an extension the Horvitz-Thompson (H-T) estimator to include not only the probability of being sampled, but also that of responding (Horvitz and Thompson, 1952; Little and Rubin, 2002). If π_i represents the probability of being selected from the population, then the general H-T estimator weights a given subject with π_i^{-1}, so that this person represents that many units in the population. The extension to include the probability of responding once selected follows similarly. A variation of the H-T method involves stratifying a sample by the levels of its predictors, then weighting according to the probability of response within each stratum (Oh and Scheuren, 1983).

More popular than the above methods is the use of propensity scores, which are an extension of stratification when the number of variables used to stratify into weighting classes becomes large and/or there are continuous predictors (Rosenbaum and Rubin, 1983, 1985; Little and Rubin, 2002). The limitation of the previous method occurs when the number of strata increases, as the number of subjects in a given stratum may be small and/or a stratum may include only non-respondents (but no respondents). In these cases, weights cannot be calculated. Additionally, continuous predictors would need to be grouped into ordinal levels in order to allow stratification. Propensity scores sidestep these issues by using logistic (or probit) regression to estimate the probability of response with all covariates as predictors. Various options exist with regard to how these probabilities are subsequently utilized for weighting. One approach is to group subjects according to ranges of probabilities, then use the average of all probabilities within a given range as the weight for subjects in that group. Alternatively, the inverse of the propensity score itself as a weight for each individual has been suggested (Cassel et al., 1983). This approach requires the data to be at least MAR.

Other methods include inverse-probability-weighted generalized estimating equations (see Section 1.3.4) (Liang and Zeger, 1986; Robins et al., 1995), post-stratification (Holt and Smith, 1979; Little, 1993; Gelman and Carlin, 2002) and raking (Ireland and Kullback, 1968; Bishop et al., 1975).

1.3.3 Imputation

Imputation is the process of filling-in missing data with means or draws from a distribution (Little and Rubin, 2002). This can be done once (*single* imputation) or repeatedly (*multiple* imputation). Once imputed, the data set is treated as com-

plete and standard statistical analyses are conducted. However, inference based on singly-imputed data usually underestimates the variation and one must adjust analyses appropriately in order to draw valid inference. Multiple imputation (MI), as discussed in Section 1.3.3.3, resolves this issue by restoring the variation in the point estimate (Rubin, 1978; Rubin, 2004). Historically, single imputation was preferred because of its ease, but the advancement of computing has made multiple imputation widely accessible in most scenarios. As alternatives to MI, one may also accurately assess the variance of point estimates by utilizing methods such as bootstrapping, jackknifing and permutations.

In general, there are *explicit* (model-based) and *implicit* imputation procedures (Little and Rubin, 2002). Explicit procedures include mean, regression and stochastic regression imputation, while implicit methods are those such as hot and cold deck imputation, and substitution.

1.3.3.1 Explicit imputation procedures In *unconditional mean imputation*, the mean of observed values of variables subject to missingness is used to fill-in missing values amongst non-respondents (Little and Rubin, 2002). This practice is quick and straightforward, but underestimates the variability of the data. Specifically, if data had actually been observed, the values would have varied across subjects, yet this procedure imputes all observations with the same quantity. If the data are stratified according to a given variable and mean imputation carried out within strata (i.e., *conditional mean imputation*), the variation is still underestimated in singly-imputed data.

Regression imputation may be used when one variable is subject to missingness and data may be assumed at least MAR (Little and Rubin, 2002). After using complete cases to regress the variable with missing values against the others, the

model is used to predict the missing observations. An extension of this idea exists in the multivariate normal setting when data are thought to be MNAR (Buck, 1960). As when imputing with means, the variation here is also underestimated, as every missing observation is imputed along the regression line, whereas the true values would actually be scattered randomly about this line.

Lastly, *stochastic* regression imputation attempts to restore the variation that is underestimated in regression imputation. This approach is an improvement over the previous methods, but is still not as accurate as multiple imputation, bootstrapping or jackknifing. The premise is to again use completers to form a regression model, but then add to it random noise (Little and Rubin, 2002). Specifically, for each subject i with missing data, the observation is imputed as $\tilde{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \eta_i$, where $\eta_i \sim N(0, s^2)$, and s^2 is the residual variance from the regression model. Little and Rubin (2002) show that in the case of bivariate, monotone, MCAR data, this method produces unbiased estimates for all parameters.

1.3.3.2 Implicit imputation procedures One example of an implicit method is *hot deck imputation*, where one draws with replacement from the respondents, fills-in the corresponding missing value, then weights that value by the number of times that respondent was drawn for imputation (Cochran, 1977). There exist alternate definitions and variations of this procedure, including hot deck within strata and the nearest-neighbor approach, which incorporates information from covariates into the random-drawing process (Rubin, 1973a; Rubin, 1973b).

Alternatively, *cold deck* imputation refers to filling in missing data with a constant value from an external data set (Little and Rubin, 2002). For example, if the same survey had been previously administered, one would use the results from that survey to impute missing values on the current form.

Lastly, *substitution* is used to replace a unit originally chosen for a sample with another unit, because the former did not respond (Little and Rubin, 2002). There is a risk of bias here, if there is some underlying difference in units who do/do not respond.

1.3.3.3 Multiple imputation, bootstrapping, jackknifing and permutations As mentioned previously, multiple imputation (MI) is a procedure that correctly restores the variation in the point estimate, so as to provide valid inference (Rubin, 1978; Rubin, 2004). However, the theory for combining test statistics and p-values over multiply-imputed data sets indicates the type-I error rate may be over- or underestimated depending on the amount of missing data, number of imputed data sets and the type-I error level (Li et al., 1991). This finding is confirmed in Section 2.3.4.

The premise of MI is to impute a given data set not once, but D times, calculating the statistics of interest each time. Then, denoting $\tilde{\theta}_d$ as the point estimate from the d^{th} imputed data set, the appropriate estimate is simply $\bar{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\tilde{\theta}_d$. The variance estimate of this value is given in Little and Rubin (2002). Inference (confidence intervals, significance tests) is drawn based on the t distribution. It may be used with explicit or implicit methods.

As an alternative to MI, one may estimate the variance of point estimates using bootstrapping or jackknifing, both of which are resampling techniques. Bootstrapping refers to drawing samples of size n with replacement from the observed data, also of size n (Efron, 1979). When missing data are present in a given bootstrap sample, values are imputed by a chosen procedure and the desired statistics calculated. As with MI, the final point estimate is the average of all bootstrapped estimates.

If the distribution of estimates is normal, a confidence interval (CI) can readily be formed based on normal theory. However, if the distribution is non-normal, one can simply use the desired percentiles to form the bootstrap CI. This method in general requires a large number of bootstrap samples, which, again, tends not to be a computational issue by current standards.

Jackknifing, the predecessor of bootstrapping, involves dropping one observation (or a set of observations) at a time from a sample, then calculating a statistic of interest. This is repeated until all observations/sets have been removed in turn, after which the jackknife estimate of the standard error of the point estimate is calculated (Miller, 1974; Efron and Gong, 1983). In the context of missing data, imputation is carried out after each data point is removed, then the procedure follows as above. In general, studies have shown the bootstrap performs better than the jackknife, but the jackknife is less computationally intensive (Efron, 1982). Little and Rubin (2002) provide the appropriate information regarding inference for point estimates after jackknifing and imputation.

For an overview of the advantages and disadvantages of MI, bootstrapping and jackknifing, and when a given procedure may be more appropriate than another, see Section 5.5 of Little and Rubin (2002).

Another data-driven method is the use of permutations, often for simulating the distribution of the test statistic under the null (Fisher, 1935; Efron, 1988; Good, 2005). This approach usually uses a subset of all possible permutations under the null, with the statistic of interest calculated in each instance. From this set of estimates, the empirical distribution is formed and the percentile of interest is used as the critical value for that set of data. This procedure is utilized in Section 2.2.5.2.

1.3.4 Model-based procedures

The three major categories of model-based procedures used in the realm of missing data are maximum likelihood, generalized estimating equations (GEE) and Bayesian methods. Because of its relevance to the subsequent studies presented in this paper, the focus of this section will be on maximum likelihood (ML) estimation and the expectation maximization (EM) algorithm. A brief discussion of GEE and Bayesian methods follows.

GEE is a procedure developed by Liang and Zeger (1986) in order to estimate the parameters of a generalized linear model when there exists a possible correlation structure due to repeated or clustered observations. The method is based on the concept of a "working correlation" (an estimate of the presumed true correlation structure), which allows the dependence between observations to be accounted for during the parameter estimation procedure. The assumption of this approach is that data are MCAR.

Robins et al. (1995) synthesized the notions of propensity scores and GEE to devise inverse-probability-weighted GEE. Specifically, logistic regression is used to predict the probability of being observed given the predictors under study. As the name implies, the inverses of these probabilities are then used to weight observations in the GEE. As a result, this method has a less strict missingness assumption, as data need only be assumed MAR.

In general, Bayesian methods consider parameters as random variables rather than constants. A "prior" distribution is specified for parameters, which is used in conjunction with information from the sampled data to create a "posterior" distribution. Details on these ideas and how Bayesian methods can be used in the missing data context can be found in Gelman et al. (1995) and Little and Rubin (2002).

1.3.4.1 The theory of maximum likelihood

Maximum likelihood as a means to estimate parameters is popular because it is (in general) easy to implement and possesses favorable properties for data that follow commonly-observed distributions such as the normal, exponential, Poisson and binomial (Pawitan, 2001). Specifically, estimates from this method are consistent and efficient (i.e., asymptotically, they obtain the Cramér-Rao lower bound). Further, ML estimation when data are missing is readily achieved using the EM algorithm (Dempster et al., 1977). The concepts of ML estimation without missing data will first be outlined, followed by the case with missing data in Section 1.3.4.2.

Within the framework of ML, the *likelihood* is a function of the parameters, $\boldsymbol{\theta}$, with the data, y, considered fixed. This contrasts a probability density function (pdf), which is a function of y for fixed $\boldsymbol{\theta}$ (Pawitan, 2001). The premise of this method is thus to solve for the value of $\boldsymbol{\theta}$ that maximizes the likelihood function for a fixed set of observed data. In other words, it asks what value of $\boldsymbol{\theta}$ is most likely given the observed data and the distribution it is assumed to follow.

It is useful to note that since the natural log is a monotonically-increasing function, maximizing the natural log of the likelihood is equivalent to maximizing the likelihood itself. This approach is often preferred because it leads to relatively straightforward optimization, given the simplified form of the likelihood after the natural log is taken. Additionally, if observations are assumed to be identically and independently distributed according to a given distribution, the joint likelihood for the sample is simply the product of n individual likelihoods (Pawitan, 2001).

Once the log-likelihood is formed (and assuming the support does not depend on the parameters), maximization is carried out by simply finding the the roots of the first derivatives with respect to the parameters (often called the *score functions*) (Pawitan, 2001). If the parameter space is bounded, boundary points should be

checked as potential maximums. In the case where the support is a function of $\boldsymbol{\theta}$, setting the derivative equal to zero is not valid, as the maximum occurs at a non-continuous endpoint on the likelihood curve. Here, the derivative is still utilized to discern whether the likelihood is an increasing or decreasing function in the parameter. Then, based on the restrictions of the parameter range, the largest or smallest value the parameter attains will maximize the likelihood. These concepts will now be formally defined for a single unknown parameter, θ, recognizing extensions exist for multiple parameters.

Denote the likelihood function by $L(\theta; y)$, where y is the observed data, θ exists in the parameter space Ω_θ and $L(\theta; y) \propto f(y; \theta)$, the pdf of Y. By definition, if $\theta \notin \Omega_\theta$, the likelihood is zero (Pawitan, 2001; Little and Rubin, 2002). Next define the (natural) log-likelihood as $\ell(\theta; y) = ln[L(\theta; y)]$ and the score function as $S(\theta) = \frac{\partial}{\partial\theta}\ell(\theta; y)$. As described above, when the support does not depend on the parameter, $\hat{\theta}_{MLE}$ is found by solving $S(\theta) \overset{set}{=} 0$ for θ. In general, $\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\, L(\theta; y) \equiv \underset{\theta}{\text{argmax}}\, \ell(\theta; y)$.

1.3.4.2 Maximum likelihood estimation with missing data When data are subject to missingness, the observed data contain both the observed outcome values, y_{obs}, and the missing data indicator, r, where R is a random variable with pdf $f(r \mid y; \psi)$. The joint distribution of Y_{obs} and R is then used to determine the *full likelihood model*:

$$L_{full}(\theta, \psi; y_{obs}, r) \propto f(y_{obs}, r; \theta, \psi) \text{ for } \theta, \psi \in \boldsymbol{\Omega}_{\theta,\psi},$$

where

$$f(y_{obs}, r; \theta, \psi) = \int f(y_{obs}, y_{mis}; \theta) f(r \mid y_{obs}, y_{mis}; \psi) dy_{mis}$$

is the joint pdf of (Y_{obs}, R) (Little and Rubin, 2002).

If data are 1) MCAR, or 2) MAR and θ and ψ are distinct ($\boldsymbol{\Omega}_{\theta,\psi} = \Omega_\theta \times \Omega_\psi$), the missing data mechanism may be ignored, implying

$$L_{full}(\theta, \psi; y_{obs}, r) = L_{ign}(\theta; y_{obs}) f(r \mid y_{obs}; \psi),$$

where $L_{ign}(\theta; y_{obs}) \propto f(y_{obs}; \theta)$ is referred to as the *ignorable likelihood* (Little and Rubin, 2002). Even though the ignorable likelihood is based on completers, its form is not always known or straightforward. From this, other approaches such as the *factored likelihood* method have arisen. See Anderson (1957) for details on this topic.

Alternatively, when data are MNAR, ψ must either be modeled jointly with θ, or the true value of ψ, ψ_0, must be assumed. Although there may exist previous knowledge to inform the choice of the model or ψ_0, if the assumption is incorrect, biased estimates of θ may result.

1.3.4.3 The EM algorithm for maximum likelihood estimation when data are subject to missingness

Because of potential computational difficulty when maximizing the full, ignorable or factored likelihoods using derivatives, a general iterative optimization method known as the *expectation maximization* (*EM*) *algorithm* has been derived (Dempster et al., 1977). Under basic conditions, the EM algorithm is guaranteed to converge to the global maximum (the maximum likelihood estimate (MLE)), although it may alternatively converge to local maxima if they exist, so that the initial state may be important depending on the shape of the curve/surface (Wu, 1983; Little and Rubin, 2002).

The guaranteed convergence is due to the ability to separate the observed-data log-likelihood, $\ell(\theta, \psi; y_{obs}, r)$, into the difference of two terms (see Section 3.1.1 for

details). The first is referred to as the Q-function, and the second, the H-function, which is guaranteed to decrease by Jensen's inequality (Dempster et al., 1977). Therefore, the difference, $Q - H$, will increase with each iteration as long as Q increases. As such, the focus of the algorithm is on maximizing Q as follows:

E-step: Given current estimates of the parameters, $\theta^{(t)}$ and $\psi^{(t)}$, calculate $Q\left[\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right]$

M-step: Maximize the Q function with respect to θ and ψ based on the expression from the E-step to obtain $\theta^{(t+1)}$ and $\psi^{(t+1)}$, and let $\theta^{(t)} = \theta^{(t+1)}$ and $\psi^{(t)} = \psi^{(t+1)}$

Given a convergence criterion, ϵ, the algorithm will eventually converge on a mode of the likelihood function, $L(\theta, \psi; y_{obs}, r)$ (Little and Rubin, 2002). The EM algorithm is preferred over methods such as Newton-Raphson because of its stability, although in general it does take longer to converge.

1.4 CONTINGENCY TABLE ANALYSIS OF DATA WITH MISSING VALUES

In the late 1960s and 1970s, missing data in the context of categorical variables was being explored through log-linear models and imputation. Specifically, Bishop and Fienberg (1969) used a log-linear model with iterative proportional fitting under the assumption of independence to estimate cell counts of a 2×2 table subject to missingness and discussed extensions to higher-dimensional tables. Blumenthal (1968) introduced the idea of missing subcategories in multinomial data and the associated bias and variance for MLEs in this case. Here, people might be completely classified

at the most general level (job title, e.g.), while lower-level classification (more specific job duties, e.g.) may be missing. Related to this, Hocking and Oxspring (1971) proposed an iterative technique to increase the precision of estimates of higher-level ("parent") categories based on partial information available in lower-levels. Hocking and Oxspring (1974) extended these methods to a contingency table setting, differentiating their work from that of Koch et al. (1972), who considered this problem from the standpoint of a linear model. Also, Chen and Fienberg (1974) discussed ML estimation for cell counts and a goodness-of-fit test in contingency tables with margins subject to missingness, addressing asymptotic variance and consistency. Fuchs (1982) combined the log-linear model approach with the EM algorithm for categorical data subject to missingness. Phillips (1993) extended the EM algorithm approach to a three-way contingency table and also considered the impact of relaxing the MAR assumption. Lastly, Lipsitz and Fitzmaurice (1996) considered a score test for independence for a general $(R \times C)$ contingency table with missing data. Compared to the likelihood ratio test, their result is easier to compute as it is not iterative.

Another relatively extensive body of literature has been established for missing categorical data in the context of survey sampling (Little, 1982; Rao and Scott, 1987). Examples include the use of log-linear models/ML estimation (Fay, 1986; Stasny, 1986; Baker and Laird, 1988) and latent models (Vermunt et al., 2008).

Of interest to the work carried out in Section 2 are the results of Gimotty and Brown (1987), which compare the empirical distribution of the chi-square goodness-of-fit test statistic after imputation to both its asymptotic distribution and that when imputation is ignored. As expected, when ignored, the test rejects too frequently, as the variation is underestimated. Also relevant are the findings of Wang (2006), where a closed-form correction factor for both a test of independence and goodness-of-fit is derived after conditional imputation of categorical variables.

1.5 A PERMUTATION-BASED CORRECTION FOR PEARSON'S CHI-SQUARE TEST ON AN IMPUTED COMPLEX OUTCOME

The second chapter in this manuscript uses the previously described notions of single imputation in the realm of binary data. Specifically, there exist contexts where it is meaningful to combine two binary outcomes, A and B, into a third binary variable, Y, referred to as a *complex outcome*. Ultimately, Pearson's chi-square test for independence between Y and another binary variable, T (treatment, e.g.), is of interest. Consider the case where A is subject to missingness and subsequently Y is as well. When data are MCAR, there exist two valid imputation procedures for the complex outcome: direct imputation of Y conditional on T, denoted $Y \,|\, T$, and indirect imputation of A conditional on B and T, denoted $A \,|\, (B, T)$. Simulation confirms single imputation based on $A \,|\, (B, T)$ is more efficient than that based on $Y \,|\, T$. In general after imputation, Pearson's test rejects the null at a rate higher than the nominal α-level, and thus correction is required. Because a closed-form solution is not clearly tractable for imputation under $A \,|\, (B, T)$, a permutation-based method is proposed. Specifically, the corrected critical value is determined by estimating the empirical distribution of the test statistic under the null. Simulation confirms this approach yields the nominal α-level under the null, and additionally that $A \,|\, (B, T)$ results in a test with higher power than $Y \,|\, T$. Additionally, the data-driven method is shown to have superior performance over multiple imputation in this context. In Section 2.4, these results are applied to neoadjuvant breast cancer clinical trial data, where a combination of drugs is compared to a single drug with regard to each treatment's effectiveness at keeping cancer from spreading into the lymph system.

1.6 A MODIFIED EM ALGORITHM FOR CONTINGENCY TABLE ANALYSIS WITH MISSING DATA

Finally, chapter 3 shows the utility of a proposed modified EM algorithm in the case where data are MNAR in the contingency table setting. The method requires no assumptions about the missing data mechanism, but does necessitate consistent initial estimates of the model parameters (obtained either through study design or from a complete, external data set). When these estimates (and possibly the external data itself) are available, the algorithm combines the information in both to yield consistent estimates potentially more efficient than those based on the external data alone. This is true even if estimates based only on the data subject to missingness would be inconsistent due to data MNAR. The basis of these results is an algebraic manipulation of the Q-function, such that most of its terms may be estimated empirically from the data. The remaining term that requires an iterative algorithm for estimation does not depend on the missing data, and thus the value or distribution of the missing data mechanism is not relevant. However, in the context of discrete data, it can be shown this approach simplifies to a special case of the general EM algorithm. The performance of the modified EM is assessed under various model structures via simulation, then applied in Section 3.4 to a data set regarding survival after surgery to treat ovarian cancer.

2.0 A PERMUTATION-BASED CORRECTION FOR PEARSON'S CHI-SQUARE TEST ON DATA WITH AN IMPUTED COMPLEX OUTCOME

2.1 INTRODUCTION

A *complex outcome*, Y, combines two or more other outcomes, providing a comprehensive summary of multiple measures. Motivation for the use of complex outcomes stems from their ability to condense a multitude of variables into a simple, workable measure while maintaining the essential information contained in the data. The variables combined and in what fashion is dictated by context.

If individual outcomes A or B (combined to form complex outcome Y) or both are subject to missingness, imputation may be conducted, then statistical analyses carried out on the imputed data set. However, indirect imputation at the A/B level affects the variation in the data differently than at the Y level, and thus analyses must be corrected appropriately depending on the method used. What follows is a description of the motivating data for this problem, a discussion of the statistical issues present, a current related method and the objectives of this paper.

To illustrate a complex outcome in practice, consider a neoadjuvant breast cancer clinical trial where researchers are interested in whether or not a novel treatment is

more successful than the current at preventing cancer from progressing into the lymph system (Robidoux et al., 2013). Neoadjuvant therapy is an alternative approach to treating cancer, where chemotherapy, hormone therapy, etc. is given *before* primary surgery to remove a tumor. This contrasts adjuvant therapy, where a tumor is resected before other treatment (radiation, chemotherapy, etc.) is prescribed. Often with the neoadjuvant approach, less tissue is removed than if surgery had been undertaken before treatment, which may result in better health and cosmetic outcomes for patients. However, since the tumor remains in the body during chemotherapy, the potential for cancer to progress into the lymph system during this time is of concern. As a result, the effectiveness of drugs at impeding this progression is of primary interest to researchers and physicians.

In order to assess the presence and extent of cancer in the lymph system after neoadjuvant therapy, lymph nodes must be removed and biopsied. Because the removal of nodes may result in lymphedema, some physicians prefer to remove only a subset of nodes, so that data is missing by design. Specifically, a tracer or dye is used to detect and remove the *sentinel nodes* (SN) − those that would be affected first if the cancer progressed. If the SN are positive for cancer, any nodes further downstream in the arm, the *axillary nodes* (AN), are removed and biopsied to assess the extent to which the cancer has spread. Conversely, the assumption given a negative SN biopsy is that all AN are also negative. In this case, no further nodes are removed and AN status is thus missing. However, in some women the SN biopsy may result in a false negative due to either incorrectly identifying the SN or a diagnostic test error. Because of this, information from women who had both their SN and AN removed is used to impute women with missing AN status. Note the presumed data structure is simplified here for illustration purposes. Specifically, it is assumed missing values of the AN depend only on SN status. In practice, however, other

covariates that may affect missingness should be considered. In this case, even if the SN were negative, physicians may decide to subsequently remove AN based on tumor size, age or weight, for example.

In clinical practice, researchers are often interested in a treatment's effect on the "overall nodal response," obtained by combining the information from all biopsies into one variable. Here, binary Y represents whether a patient had either no cancer in any lymph nodes or at least one node with cancer. In other words, Y is a complex outcome formed by combining information from the AN (A) and SN (B).

To determine whether or not a novel treatment results in a lower rate of cancer in the lymph nodes, Pearson's chi-square test for independence may be used. Specifically, independence between Y and T is tested, where T is a treatment indicator. However, when missing AN status is imputed, the test performed on singly-imputed data will underestimate the variation due to imputation and reject the null more often than it should. Because of this, correction is needed for inference to be meaningful.

For single imputation using $Y \,|\, T$, which represents imputing Y (missing if A was missing) conditionally on T, Wang (2006) developed a closed-form correction factor for Pearson's test, which adjusted the observed test statistic based on the percentage of missing data in the sample. However, this imputation scheme is naïve compared to a single imputation of A given B and T (before the calculation of Y), denoted $A \,|\, (B, T)$, which utilizes more information in the data. Note in this case $A \,|\, (B, T)$ is equivalent to $Y \,|\, (B, T)$.

However, extending Wang's (2006) result to this more complicated setting is not necessarily viable. Specifically, the assumption under the null that $Y \perp T$ is imperative to developing the closed-form correction factor. In contrast, imputation using $A \,|\, (B, T)$ still assumes under the null that $(A, B) \perp T \Rightarrow A \perp T$ and $B \perp T$, but not that $A \perp B$. Because of this, the asymptotic properties established by Wang

(2006) do not hold for $A \,|\, (B, T)$. As a practical and valid alternative, a permutation-based approach that estimates the distribution of the test statistic under the null is proposed (Fisher, 1935; Efron, 1988; Good, 2005). From this distribution, the test statistic that results in the expected type-I error rate is selected as the appropriate adjusted critical value.

The goals of this paper are to: 1) identify which imputation procedures are and are not valid in the context of a complex outcome, and quantify via simulation the level of bias and variation in each procedure; 2) illustrate with simulation that the inflated rate of rejection of Pearson's test is not equivalent for all valid imputation procedures, and that inflation due to $A \,|\, (B, T)$ depends on the percent of missing data as well as the distribution of (A, B); 3) describe a permutation-based empirical method to correct Pearson's test given imputation under $A \,|\, (B, T)$, and show it results in higher power than $Y \,|\, T$ with simulation; and 4) use simulation to show the proposed method in (3) is more successful at achieving the nominal type-I error rate than multiple imputation in this context. Aims (1) and (2) additionally show imputation using $Y \,|\, T$ is less efficient than $A \,|\, (B, T)$.

2.2 METHODS

In the analysis of the above clinical trial data, the goal is to estimate the response rate of Y (overall nodal status) across levels of T (treatment), and conduct a test of independence between these two variables. Because of potentially missing AN status, imputation is utilized, and the chosen procedure should yield consistent point estimates of the cell probabilities of the multinomial distribution defined by Y and T in order to be considered valid.

2.2.1 Notation for discrete data with missing values

In order to remain congruous with Wang (2006), much of the notation used in this paper is the same or similar, and is extended or altered when needed.

Consider a joint outcome vector, $\mathbf{F} = (A, B)'$, where A and $B \in \{1, 2\}$ so that $(A, B) \in \{(1, 1), (1, 2), (2, 1), (2, 2)\}$. Without missing data, $\mathbf{F}_{ij}$ jointly form a multinomial random variable, with each $\mathbf{F}_{ij}$ representing the number of times $(A, B)' = (i, j)'$ is observed over n trials. Let A be subject to missingness while B is fully-observed.

In general, a *complex outcome* is a function of two or more other outcomes. Here, define the binary complex outcome $Y \in \{1, 2\}$, derived from $\mathbf{F} = (A, B)'$ and indexed by k, as

$$Y = \begin{cases} 2 & \text{if } A = B = 2 \\ 1 & \text{o.w.} \end{cases} . \tag{2.1}$$

Let $T \in \{1, 2\}$ be a (fully-observed) treatment indicator indexed by l, where

$$T = \begin{cases} 2 & \text{if subject is in the treatment group} \\ 1 & \text{if subject is in the control group} \end{cases} \tag{2.2}$$

and define the following for $i, j, k, l \in \{1, 2\}$:

$$\begin{aligned} & p_{ijl} = P[(A, B, T) = (i, j, l)] \\ & p_{ij\cdot} = P[(A, B) = (i, j)], \text{ and similar for other probabilities} \\ & p_{A=i \,|\, B=j} = P(A = i \,|\, B = j) = \frac{p_{ij\cdot}}{p_{\cdot j\cdot}}, \text{ and similar for other probabilities} \\ & \mathbf{p} = (p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221}, p_{222})' \\ & q_{kl} = P[(Y, T) = (k, l)] \\ & \mathbf{q} = (q_{11}, q_{12}, q_{21}, q_{22})'. \end{aligned} \tag{2.3}$$

If observations are indexed by $m \in \{1, ..., n\}$, then

C_B = subset of $\{1, \ldots, n\}$ where B is observed, but A is missing

C_C = subset of $\{1, \ldots, n\}$ where$A$ and B are both observed (i.e., "completers")

$n^B = \sum_m I\{m \in C_B\}$

$n^C = \sum_m I\{m \in C_C\}$

$n^C_{ijl} = \sum_{m \in C_C} I\{(A, B, T)_m = (i, j, l)\}$

$n^*_{\cdot jl} = \sum_{m \in C_*} I\{(B, T)_m = (j, l)\}$, where * is B or C; similar for other counts

$\pi_B = \frac{n^B}{n}$ = probability A is missing while B is observed

$\pi_C = \frac{n^C}{n}$ = probability both A and B are observed.

2.2.2 Assumptions

1. Observations are independent of one another
2. Data are missing completely at random (see Section 2.4 for an exception where data are missing at random)
3. Missingness of A depends only on B and no other potential covariates
4. Single imputation is carried out with simple random sampling under the designated imputation scheme

2.2.3 Validity of imputation procedures

Based on Wang (2006) and as given in (2.3), C_B is the set of indices for which B is observed and A is missing. Once A is imputed, estimates based on the imputed data from those n^B people are given by $\hat{p}^B_{ijl} = \frac{1}{n^B} \sum_{m \in C_B} I\{(A, B, T)_m = (i, j, l)\}$ and

$\hat{q}_{kl}^{B} = \frac{1}{n^B} \sum_{m \in C_B} I\{(Y,T)_m = (k,l)\}$. $\hat{p}_{ijl}^{C}$ and $\hat{q}_{kl}^{C}$ are defined similarly for n^C and $m \in C_C$, except they are based on the fully-observed data. Equation (2.4) gives the expression for estimates of the parameters of the multinomial distribution (A, B, T) based on the imputed data set:

$$\hat{p}_{ijl}^{I} = \frac{n^B \hat{p}_{ijl}^{B} + n^C \hat{p}_{ijl}^{C}}{n}. \tag{2.4}$$

Subsequently, let $\hat{\mathbf{p}}^* = (\hat{p}^*_{111}, \hat{p}^*_{112}, \hat{p}^*_{121}, \hat{p}^*_{122}, \hat{p}^*_{211}, \hat{p}^*_{212}, \hat{p}^*_{221}, \hat{p}^*_{222})'$, where * can be B, C or I, and similar for $\hat{\mathbf{q}}^*$.

In general, it should be true that $\hat{p}^I_{ijl} \xrightarrow{p} p_{ijl}$ or $\hat{q}^I_{kl} \xrightarrow{p} q_{kl}\ \forall\ i,j,k,l$ as $n \to \infty$, or the point estimate is not consistent (Casella and Berger, 2002). Since $\hat{\mathbf{p}}^I$ is a weighted function of $\hat{\mathbf{p}}^B$ and $\hat{\mathbf{p}}^C$ (Eq. (2.4)), and $\hat{\mathbf{p}}^C$ is consistent under the assumption of MCAR, the consistency of $\hat{\mathbf{p}}^B$ (or $\hat{\mathbf{q}}^B$) is of interest.

For simplicity, consider without loss of generality the element $\hat{p}^B_{222}$ of $\hat{\mathbf{p}}^B$ instead of the entire $\hat{\mathbf{p}}^B$ vector, or $\hat{q}^B_{22}$ instead of $\hat{\mathbf{q}}^B$, recognizing the following results apply analogously to all elements of the vector. Let $\hat{n}^B_{222}$ be the number of subjects in C_B for which $A = B = T = 2$ after imputation of A. Note $Y = 2$ when $A = B = 2$, so that $\hat{n}^B_{222}$ is equivalent to the estimate of the number of subjects in C_B for which $Y = T = 2$ after imputation of Y.

2.2.3.1 Marginal imputation of a complex outcome ($\boldsymbol{Y}$)

In this scenario, imputation is carried out by first calculating $\hat{q}^C_{2\cdot}$, then sampling from $Y_m \sim Bernoulli\left(\hat{q}^C_{2\cdot}\right)$ for $m \in C_B$, where $Y_m \in \{1,2\}$. Subsequently, $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot\cdot 2}, \hat{q}^C_{2\cdot}\right)$. Define $o(1)$ as a random variable such that $\lim_{n\to\infty} o(1) \xrightarrow{a.s.} 0$. It follows that

$$\hat{q}^B_{22} = \frac{\hat{n}^B_{222}}{n^B}$$

$$
\begin{aligned}
&= \frac{n^B_{\cdot\cdot 2}}{n^B} \frac{\hat{n}^B_{222}}{n^B_{\cdot\cdot 2}} \\
&= \hat{q}^B_{\cdot 2} \left[\hat{q}^C_{2\cdot} + o(1) \right] \text{ as } n^B_{\cdot\cdot 2} \to \infty \qquad (2.5) \\
&\xrightarrow{p} q_{\cdot 2} q_{2\cdot} \text{ as } n \to \infty \qquad (2.6) \\
&= q_{22} \text{ iff } T \perp Y.
\end{aligned}
$$

Equation (2.5) follows from the fact that $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot\cdot 2}, \hat{q}^C_{2\cdot}\right)$ and the strong law of large numbers (SLLN): In general, if $X_1, X_2, ..., X_n$ represents a sequence of random variables, then the SLLN states $\bar{X}_n \xrightarrow{a.s.} \mu$ as $n \to \infty$, where $\bar{X}_n$ and μ are the sample and population means, respectively (Casella and Berger, 2002). Since a proportion is a special case of a mean, the SLLN applies here as well and indicates $\frac{\hat{n}^B_{222}}{n^B_{\cdot\cdot 2}} = \hat{q}^C_{2\cdot} + o(1)$ as $n^B_{\cdot\cdot 2} \to \infty$. Equation (2.6) follows by first noting $\hat{q}^B_{kl} \xrightarrow{a.s.} q_{kl}$ and $\hat{q}^C_{k'l'} + o(1) \xrightarrow{a.s.} q_{k'l'}$ as $n \to \infty$ by the SLLN, where k, l may or may not be equal to k', l'. Then, since almost sure convergence implies convergence in probability (Rohatgi, 1976), $\hat{q}^B_{kl} \left[\hat{q}^C_{k'l'} + o(1) \right] \xrightarrow{p} q_{kl} q_{k'l'}$ as $n \to \infty$ (Bain and Engelhardt, 1992).

Since in general the complex outcome, Y, cannot be assumed to be independent of treatment, marginal imputation of Y is not valid.

2.2.3.2 Marginal imputation of the missing component of a complex outcome (A)

Here, A is first imputed, then Y is calculated based on the imputed data set. Imputation of A is achieved by sampling from $A_m \sim Bernoulli\left(\hat{p}^C_{2\cdot\cdot}\right)$ for $m \in C_B$, so that $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot 22}, \hat{p}^C_{2\cdot\cdot}\right)$. Then,

$$
\begin{aligned}
\hat{p}^B_{222} &= \frac{\hat{n}^B_{222}}{n^B} \\
&= \frac{n^B_{\cdot 22}}{n^B} \frac{\hat{n}^B_{222}}{n^B_{\cdot 22}} \\
&= \hat{p}^B_{\cdot 22} \left[\hat{p}^C_{2\cdot\cdot} + o(1) \right] \text{ as } n^B_{\cdot 22} \to \infty
\end{aligned}
$$

$$\xrightarrow{p} p_{\cdot 22} p_{2\cdot\cdot} \text{ as } n \to \infty$$
$$= p_{222} \text{ iff } (B, T) \perp A \Rightarrow \text{iff } B \perp A \text{ and } T \perp A.$$

In general, it will not be true that $B \perp A$ and $T \perp A$, so this imputation procedure is also invalid.

2.2.3.3 Conditional imputation of the missing component of a complex outcome (A) given the other component (B) Here, the information from B is used to impute A, sampling randomly from $(A \,|\, B = 1)_m \sim Bernoulli\left(\hat{p}^C_{A=2\,|\,B=1}\right)$ if $m \in C_B$ and $B_m = 1$, or $(A \,|\, B = 2)_m \sim Bernoulli\left(\hat{p}^C_{A=2\,|\,B=2}\right)$ if $m \in C_B$ and $B_m = 2$. Subsequently, $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot 22}, \hat{p}^C_{A=2\,|\,B=2}\right)$. Then,

$$\begin{aligned}
\hat{p}^B_{222} &= \frac{\hat{n}^B_{222}}{n^B} \\
&= \frac{n^B_{\cdot 22}}{n^B} \frac{\hat{n}^B_{222}}{n^B_{\cdot 22}} \\
&= \hat{p}^B_{\cdot 22} \left[\hat{p}^C_{A=2\,|\,B=2} + o(1)\right] \text{ as } n^B_{\cdot 22} \to \infty \\
&\xrightarrow{p} P(B = T = 2) P(A = 2 \,|\, B = 2) \text{ as } n \to \infty \\
&= P(T = 2 \,|\, B = 2) P(B = 2) P(A = 2 \,|\, B = 2) \\
&= P(A = T = 2 \,|\, B = 2) P(B = 2) \text{ iff } (T \perp A) \,|\, B \\
&= p_{222} \text{ iff } (T \perp A) \,|\, B.
\end{aligned}$$

Similar to the previous two cases, $(T \perp A) \,|\, B$ is not a reasonable assumption, so that this imputation procedure is not valid.

2.2.3.4 Conditional imputation of the missing component of a complex outcome (A) given a covariate (T) Imputing A based on treatment means

sampling from $(A \,|\, T = 1)_m \sim Bernoulli\left(\hat{p}^C_{A=2\,|\,T=1}\right)$ if $m \in C_B$ and $T_m = 1$, or $(A \,|\, T = 2)_m \sim Bernoulli\left(\hat{p}^C_{A=2\,|\,T=2}\right)$ if $m \in C_B$ and $T_m = 2$, so that $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot 22}, \hat{p}^C_{A=2\,|\,T=2}\right)$. It follows that

$$\begin{aligned}
\hat{p}^B_{222} &= \frac{\hat{n}^B_{222}}{n^B} \\
&= \frac{n^B_{\cdot 22}}{n^B}\frac{\hat{n}^B_{222}}{n^B_{\cdot 22}} \\
&= \hat{p}^B_{\cdot 22}\left[\hat{p}^C_{A=2\,|\,T=2} + o(1)\right] \text{ as } n^B_{\cdot 22} \to \infty \\
&\xrightarrow{p} P(B = T = 2)P(A = 2 \,|\, T = 2) \text{ as } n \to \infty \\
&= P(B = 2 \,|\, T = 2)P(T = 2)P(A = 2 \,|\, T = 2) \\
&= P(A = B = 2 \,|\, T = 2)P(T = 2) \text{ iff } (A \perp B) \,|\, T \\
&= p_{222} \text{ iff } (A \perp B) \,|\, T.
\end{aligned}$$

In general it is not true that $(A \perp B) \,|\, T$, as the motivation for forming a complex outcome based on A and B is from the belief they are somehow related. As such, this is not a reasonable imputation procedure.

2.2.3.5 Conditional imputation of a complex outcome (Y) given a covariate (T) In contrast to imputing A given treatment, Y given T results in a consistent estimate of q_{22}. Here, imputation is carried out as $(Y \,|\, T = 1)_m \sim Bernoulli\left(\hat{q}^C_{Y=2\,|\,T=1}\right)$ if $m \in C_B$ and $T_m = 1$, or $(Y \,|\, T = 2)_m \sim Bernoulli\left(\hat{q}^C_{Y=2\,|\,T=2}\right)$ if $m \in C_B$ and $T_m = 2$. Then, $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot\cdot 2}, \hat{q}^C_{Y=2\,|\,T=2}\right)$ and

$$\begin{aligned}
\hat{q}^B_{22} &= \frac{\hat{n}^B_{222}}{n^B} \\
&= \frac{n^B_{\cdot\cdot 2}}{n^B}\frac{\hat{n}^B_{222}}{n^B_{\cdot\cdot 2}}
\end{aligned}$$

$$
\begin{aligned}
&= \hat{q}^B_{\cdot 2}\left[\hat{q}^C_{Y=2\,|\,T=2} + o(1)\right] \text{ as } n^B_{\cdot\cdot 2} \to \infty \\
&\xrightarrow{p} q_{\cdot 2} q_{Y=2\,|\,T=2} \text{ as } n \to \infty \\
&= q_{22}.
\end{aligned}
$$

This indicates that under this imputation procedure, $\hat{q}^B_{kl}$ is a consistent estimator of q_{kl}.

2.2.3.6 Conditional imputation of the missing component of a complex outcome (A) given the other component (B) and a covariate (T) A imputed conditionally on B and T is the only other consistent estimate of p_{222} (q_{22}) in this context. When $m \in C_B$ and $B_m = T_m = 1$, sampling is done from $(A\,|\,B = T = 1)_m \sim Bernoulli\left(\hat{p}^C_{A=2\,|\,B=T=1}\right)$, and similarly for other combinations of B_m and T_m. This implies $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot 22},\, \hat{p}^C_{A=2\,|\,B=T=2}\right)$. Then,

$$
\begin{aligned}
\hat{p}^B_{222} &= \frac{\hat{n}^B_{222}}{n^B} \\
&= \frac{n^B_{\cdot 22}}{n^B}\frac{\hat{n}^B_{222}}{n^B_{\cdot 22}} \\
&= \hat{p}^B_{\cdot 22}\left[\hat{p}^C_{A=2\,|\,B=2,T=2} + o(1)\right] \text{ as } n^B_{\cdot 22} \to \infty \\
&\xrightarrow{p} P(B = T = 2)P(A = 2\,|\,B = T = 2) \text{ as } n \to \infty \\
&= p_{222}.
\end{aligned}
$$

Thus, as in Section 2.2.3.5, in general for this type of imputation $\hat{p}^B_{ijl}$ is a consistent estimator of p_{ijl}.

2.2.4 Asymptotic distributions of consistent estimators under the null of independence

As shown in Section 2.2.3, the only imputation procedures that result in a consistent estimator of $\mathbf{p}$ without any independence assumptions are $Y \,|\, T$ and $A \,|\, (B, T)$ (equivalent to $Y \,|\, (B, T)$), and thus the distributional properties in these two cases under the null of $A \perp T$ and $B \perp T \Rightarrow Y \perp T$ are of interest. Note that independence is not assumed between A and B, as the motivation for forming a complex outcome based on A and B is that they are indeed dependent.

Section 2.2.4.1 shows that in the case of $Y \,|\, T$, the covariance of the distribution of interest is analogous to that given by Wang (2006). However, Section 2.2.4.2 indicates imputation under $A \,|\, (B, T)$ does not allow for a clearly tractable derivation of the covariance matrix.

2.2.4.1 Conditional imputation of a complex outcome given a covariate ($\boldsymbol{Y \,|\, T}$)

The asymptotic results in this situation are equivalent to those established by Wang (2006). Given the definitions in (2.3), $q_{11} = p_{111} + p_{121} + p_{211}$, $q_{12} = p_{112} + p_{122} + p_{212}$, $q_{21} = p_{221}$ and $q_{22} = p_{222}$. Thus, Y is analogous to Wang's (2006) variable A, while T is equivalent to Wang's B, where A and $B \in \{1, 2\}$ and A is subject to missingness while B is fully-observed (since treatment is always observed here).

Assuming Y and T are independent, $\pi_C > 0$ and conditional imputation using $Y \,|\, T$, Wang's (2006) Theorem 1 indicates $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^q)$, where

$\hat{\mathbf{q}}^I$ is defined analogously to $\hat{\mathbf{p}}^I$

$\boldsymbol{\Sigma}^q = (\pi_C^{-1} + 1 - \pi_C)(\mathbf{Q}_Y \otimes \mathbf{Q}_T) + \frac{\pi_C + 2\pi_C\pi_T + \pi_T^2}{\pi_C}\left[\mathbf{Q}_Y \otimes (\mathbf{q}_T\mathbf{q}_T')\right] + \left[(\mathbf{q}_Y\mathbf{q}_Y') \otimes \mathbf{Q}_T\right]$

$\otimes$ represents the Kronecker product

$\mathbf{Q}_Y = diag\{\mathbf{q}_Y\} - \mathbf{q}_Y\mathbf{q}_Y'$, where $\mathbf{q}_Y = (q_{1\cdot}, q_{2\cdot})'$

$\mathbf{Q}_T = diag\{\mathbf{q}_T\} - \mathbf{q}_T\mathbf{q}_T'$, where $\mathbf{q}_T = (q_{\cdot 1}, q_{\cdot 2})'$

π_T = probability that T is observed and Y is missing.

The closed form of $\mathbf{\Sigma}^q$ hinges, in part, on the fact that $E[\hat{\mathbf{q}}^B \,|\, \sigma(C)] - \mathbf{q}$ may be written as

$$E[\hat{\mathbf{q}}^B \,|\, \sigma(C)] - \mathbf{q}_Y \otimes \mathbf{q}_T \tag{2.7}$$

since $Y \perp T$ under the null, and where $\sigma(C)$ represents the set of observed data: $\{n^B$, n^C, $(A, B, T)_m$ when $m \in C_C$ and $(B, T)_m$ when $m \in C_B\}$. Again for simplification, consider the element q_{22} from the $\mathbf{q}$ vector and note $E\left[\hat{q}^B_{22} \,|\, \sigma(C)\right]$ is constant given $\sigma(C)$, so that

$$\begin{aligned} E\left[\hat{q}^B_{22} \,|\, \sigma(C)\right] &= E\left[\frac{\hat{n}^B_{222}}{n^B} \,|\, \sigma(C)\right] \\ &= \frac{1}{n^B} n^B_{\cdot\cdot 2} \hat{q}^C_{Y=2\,|\,T=2} \\ &= \hat{q}^B_{\cdot 2} \frac{\hat{q}^C_{22}}{\hat{q}^C_{\cdot 2}}. \end{aligned} \tag{2.8}$$

Equation (2.8) comes from the fact that $\hat{n}^B_{222} \sim BIN\left(n^B_{\cdot\cdot 2}, \hat{q}^C_{Y=2\,|\,T=2}\right)$ as defined in Section 2.2.3.5. Then, given $Y \perp T$ under the null,

$$\begin{aligned} E\left[\hat{q}^B_{22} \,|\, \sigma(C)\right] - q_{2\cdot}q_{\cdot 2} &= \left(\hat{q}^B_{\cdot 2}\frac{\hat{q}^C_{22}}{\hat{q}^C_{\cdot 2}} - q_{2\cdot}q_{\cdot 2}\right) \\ &= \left(q_{\cdot 2}\frac{\hat{q}^C_{22}}{\hat{q}^C_{\cdot 2}} - q_{2\cdot}q_{\cdot 2}\right) + o(1) \;\; \because \;\; \hat{q}^B_{kl} \xrightarrow{a.s.} q_{kl} \text{ by the SLLN} \\ &= \left(q_{\cdot 2}\frac{\hat{q}^C_{22}}{q_{\cdot 2} - q_{\cdot 2} + \hat{q}^C_{\cdot 2}} - q_{2\cdot}q_{\cdot 2}\right) + o(1) \\ &= \left[q_{\cdot 2}\frac{\hat{q}^C_{22}}{q_{\cdot 2} + (\hat{q}^C_{\cdot 2} - q_{\cdot 2})} - q_{2\cdot}q_{\cdot 2}\right] + o(1) \end{aligned}$$

$$
\begin{aligned}
&= \left\{ \frac{\hat{q}^C_{22}}{\left[\frac{q_{\cdot 2}+\left(\hat{q}^C_{\cdot 2}-q_{\cdot 2}\right)}{q_{\cdot 2}}\right]} - q_{2\cdot}q_{\cdot 2} \right\} + o(1) \\
&= \left(\frac{\hat{q}^C_{22}}{1+\frac{\hat{q}^C_{\cdot 2}-q_{\cdot 2}}{q_{\cdot 2}}} - q_{2\cdot}q_{\cdot 2} \right) + o(1) \\
&= \left[\hat{q}^C_{22}\left(1-\frac{\hat{q}^C_{\cdot 2}-q_{\cdot 2}}{q_{\cdot 2}}\right) - q_{2\cdot}q_{\cdot 2} \right] + o(1) \\
&\quad \text{by the Maclaurin series, } \frac{1}{1-(-x)} = 1 - x + x^2 - x^3 + \cdots \\
&= \left[\hat{q}^C_{22} - \frac{\hat{q}^C_{22}\left(\hat{q}^C_{\cdot 2}-q_{\cdot 2}\right)}{q_{\cdot 2}} - q_{2\cdot}q_{\cdot 2} \right] + o(1) \\
&= \left(\hat{q}^C_{22} - q_{2\cdot}q_{\cdot 2}\right) - \frac{\hat{q}^C_{22}}{q_{\cdot 2}}\left(\hat{q}^C_{\cdot 2}-q_{\cdot 2}\right) + o(1) \\
&= \left(\hat{q}^C_{22} - q_{2\cdot}q_{\cdot 2}\right) - \frac{q_{22}}{q_{\cdot 2}}\left(\hat{q}^C_{\cdot 2}-q_{\cdot 2}\right) + o(1) \\
&\quad \text{by Slutsky's Theorem} \\
&= \left(\hat{q}^C_{22} - q_{2\cdot}q_{\cdot 2}\right) - \frac{q_{2\cdot}q_{\cdot 2}}{q_{\cdot 2}}\left(\hat{q}^C_{\cdot 2}-q_{\cdot 2}\right) + o(1) \ \because\ Y \perp T \\
&= \left(\hat{q}^C_{22} - q_{2\cdot}q_{\cdot 2}\right) - q_{2\cdot}\left(\hat{q}^C_{\cdot 2}-q_{\cdot 2}\right) + o(1). \qquad (2.9)
\end{aligned}
$$

Based on the simplification shown in (2.9), (2.7) may be rewritten as a linear function of $\mathbf{q}_Y$, $\mathbf{q}_T$ and indicator vectors for Y and T. Specifically, define $\mathbf{I}^Y_m$ as a two-dimensional vector with its first element equal to 1 if $Y = 1$, 0 otherwise; and the second element equal to 1 if $Y = 2$, 0 otherwise; and similarly for $\mathbf{I}^T_m$. For each of the n observations, these vectors will be independent since $Y \perp T$. Note $\frac{1}{n^C}\sum_{m \in C_C} \mathbf{I}^Y_m = \hat{\mathbf{q}}^C_Y$ and $\frac{1}{n^C}\sum_{m \in C_C} \mathbf{I}^T_m = \hat{\mathbf{q}}^C_T$, so that (2.7) becomes

$$
E[\hat{\mathbf{q}}^B \,|\, \sigma(C)] - \mathbf{q}_Y \otimes \mathbf{q}_T = \hat{\mathbf{q}}^C - \mathbf{q}_Y \otimes \mathbf{q}_T - \mathbf{q}_Y \otimes \hat{\mathbf{q}}^C_T + \mathbf{q}_Y \otimes \mathbf{q}_T + o(1)
$$

$$
\begin{aligned}
&= \frac{1}{n^C} \sum_{m \in C_C} \mathbf{I}_m^Y \otimes \mathbf{I}_m^T - \mathbf{q}_Y \otimes \mathbf{q}_T - \frac{1}{n^C} \sum_{m \in C_C} \mathbf{q}_Y \otimes \mathbf{I}_m^T + \\
&\quad \mathbf{q}_Y \otimes \mathbf{q}_T + o(1) \\
&= \frac{1}{n^C} \sum_{m \in C_C} [(\mathbf{I}_m^Y - \mathbf{q}_Y) \otimes (\mathbf{I}_m^T - \mathbf{q}_T) + (\mathbf{I}_m^Y - \mathbf{q}_Y) \otimes \mathbf{q}_T] + \\
&\quad o(1).
\end{aligned}
$$

Based on the above results, Wang (2006) proceeds to derive the variance and covariance of all terms in (2.7), thus determining the distribution of $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)$. Subsequently, the distribution of Pearson's test statistic is derived as a function of this distribution, leading to the closed-form correction factor (given in Section 2.2.5.1). The final form of (2.7) given above is imperative for the simplification required to derive the covariance matrix of $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)$ and subsequent distribution of the test statistic. In Section 2.2.4.2, it is noted the assumption of $Y \,|\, T$ used by Wang (2006) for such simplification is violated in the case of imputation under $A \,|\, (B, T)$, so that Wang's results do not apply analogously to this situation. See Wang (2006) for further detail.

2.2.4.2 Conditional imputation of the missing component of a complex outcome given the other component and a covariate ($\boldsymbol{A \,|\, (B, T)}$) Since here A is imputed based on B and T, $\mathbf{p}$, not $\mathbf{q}$, must be considered. Specifically, the goal is to derive the distribution of $\sqrt{n}\left(\hat{\mathbf{p}}^I - \mathbf{p}\right)$ after imputation similarly to the derivation for $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)$. However, since Pearson's chi-square test is between Y and T, the test is carried out at the level of $\mathbf{q}$ and thus the distribution of $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)$ must be determined as a function of the distribution of $\sqrt{n}\left(\hat{\mathbf{p}}^I - \mathbf{p}\right)$, with the ultimate goal of finding the distribution of the test statistic based on the distribution of $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)$.

Let $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)_{AB}$ represent the asymptotic distribution of $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)$ after imputation based on $A \,|\, (B, T)$, so as to distinguish it from $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)$ after imputation via $Y \,|\, T$. As in Section 2.2.4.1, $\mathbf{q}$ is a linear mapping of $\mathbf{p}$, so that the asymptotic distribution of interest is attained by mapping $\sqrt{n}\left(\hat{\mathbf{p}}^I - \mathbf{p}\right)$ to $\sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)_{AB}$ via

$$\sqrt{n}\left(\hat{\mathbf{p}}^I - \mathbf{p}\right) = \sqrt{n}\begin{pmatrix} \hat{p}^I_{111} - p_{111} \\ \hat{p}^I_{112} - p_{112} \\ \hat{p}^I_{121} - p_{121} \\ \hat{p}^I_{122} - p_{122} \\ \hat{p}^I_{211} - p_{211} \\ \hat{p}^I_{212} - p_{212} \\ \hat{p}^I_{221} - p_{221} \\ \hat{p}^I_{222} - p_{222} \end{pmatrix}$$

$$\Rightarrow \sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)_{AB} = \sqrt{n}\begin{pmatrix} \hat{q}^I_{11} - q_{11} \\ \hat{q}^I_{12} - q_{12} \\ \hat{q}^I_{21} - q_{21} \\ \hat{q}^I_{22} - q_{22} \end{pmatrix} = \sqrt{n}\begin{pmatrix} \left(\hat{p}^I_{111} - p_{111}\right) + \left(\hat{p}^I_{121} - p_{121}\right) + \left(\hat{p}^I_{211} - p_{211}\right) \\ \left(\hat{p}^I_{112} - p_{112}\right) + \left(\hat{p}^I_{122} - p_{122}\right) + \left(\hat{p}^I_{212} - p_{212}\right) \\ \hat{p}^I_{221} - p_{221} \\ \hat{p}^I_{222} - p_{222} \end{pmatrix}$$

$$\Rightarrow \sqrt{n}\left(\hat{\mathbf{q}}^I - \mathbf{q}\right)_{AB} = \sqrt{n}\mathbf{H}\left(\hat{\mathbf{p}}^I - \mathbf{p}\right), \text{ where } \mathbf{H} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The above result depends on the attainment of the closed-form distribution of $\sqrt{n}\left(\hat{\mathbf{p}}^I - \mathbf{p}\right)$. However, this is complicated by the fact that under the null, $(A, B) \perp T \Rightarrow A \perp T$ and $B \perp T$, but not that $A \perp B$, yet imputation of A is conditional

on B. This contrasts $Y \,|\, T$, where Y is imputed conditionally on T, and Y and T are independent under the null. Also, $\mathbf{p}$ is three-dimensional (meaning it is a function of three variables), while $\mathbf{q}$ is two-dimensional. Although one maps from a 3-D to 2-D space with a constant matrix, $\mathbf{H}$, further simplifications are still done using A, B and T (i.e. one still works with $\mathbf{p}$). This higher-dimensional problem, in addition to the dependence between A and B, indicates deriving the distribution of the test statistic under the null is potentially not feasible.

For a practical solution to correcting the rate of rejection of Pearson's test in this setting, a permutation-based method is proposed, in contrast to deducing the closed-form correction factor. As discussed in detail in Section 2.3, there is motivation to use imputation of $A \,|\, (B, T)$ over $Y \,|\, T$ despite a lack of a closed-form adjustment, as it is more efficient and powerful.

2.2.5 Correction to Pearson's χ^2 test of independence under valid imputation

With no missing data, Pearson's χ^2 test statistic between Y and T is given by

$$X_Y^2 := n \sum_{kl} \frac{(\hat{q}_{kl} - \hat{q}_{k\cdot}\hat{q}_{\cdot l})^2}{\hat{q}_{k\cdot}\hat{q}_{\cdot l}} \xrightarrow{d} \chi^2_{(y-1)(t-1)}, \tag{2.10}$$

where y and t are the number of categories of Y and T, respectively. However, if there are missing observations, imputation of one or both variables affects the distribution of the test statistic under the null. Specifically, imputation adds another level of variation to the data, so that standard inference without correction will undermine the true variation in the data. In Section 2.2.3, correction factors for the two valid imputation procedures are established.

2.2.5.1 Conditional imputation of a complex outcome given a covariate ($\boldsymbol{Y} \mid \boldsymbol{T}$)

When imputation is done according to $Y \mid T$, a closed-form solution for the correction factor exists, as deduced by Wang (2006): Given the asymptotic variance found in Section 2.2.4.1, Wang (2006) showed $(\pi_C^{-1} + 1 - \pi_C)^{-1}$ is the appropriate correction. In other words, under the null, $\frac{X_Y^2}{\pi_C^{-1}+1-\pi_C} \xrightarrow{d} \chi^2_{(y-1)(t-1)} = \chi^2_1$ (since Y and T are both binary in this case). Note this expression depends only on the proportion of completers in the sample.

2.2.5.2 Conditional imputation of the missing component of a complex outcome given the other component and a covariate ($\boldsymbol{A} \mid (\boldsymbol{B}, \boldsymbol{T})$)

As discussed in Section 2.2.4.2, the form of the variance-covariance under $A \mid (B, T)$ is not necessarily achievable, and thus an algorithm that utilizes permutations is proposed in order to obtain the adjusted critical value. Specifically, *Algorithm 1* outlines the procedure used to construct the empirical distribution of the test statistic under the null of $(A, B) \perp T$, which then provides an estimate of the adjusted critical value (Fisher, 1935; Efron, 1988; Good, 2005).

Algorithm 1: Determine the adjusted critical value for Pearson's chi-square test after imputation under $A \mid (B, T)$ based on the empirical distribution of the test statistic under the null:

Step 1: For an observed sample of size n from a multinomial distribution with data assumed to be MCAR, impute missing observations with valid procedure $A \mid (B, T)$ based on complete cases as shown in Section 2.2.3.6, then calculate Y.

Step 2: Conduct Pearson's test for independence between Y and T and calculate the test statistic, S.

Step 3: For $d = 1, \ldots, D$:

(a) Permute T randomly in the original data to simulate the null of $(A, B) \perp T$.

(b) Impute missing values of A using $A \,|\, (B, T)$ and calculate Y appropriately.

(c) Conduct Pearson's test between Y and T and record $S^{(d)}$, the observed test statistic from the d^{th} permuted and imputed data set.

Step 4: Organize the set of D test statistics from *Step 3*, $\left\{S^{(1)}, S^{(2)}, \ldots, S^{(D)}\right\}$, in ascending order and find $S^*_{1-\alpha}$, the $(1-\alpha)^{\text{th}}$ percentile of the empirical distribution of the test statistic, which is the corrected critical value of interest.

Step 5: Reject the null of $Y \perp T$ if $S > S^*_{1-\alpha}$.

2.2.5.3 Comparison of permutation-based method to multiple imputation Given multiple imputation (MI) is a procedure used to adjust inference on point estimates after imputation, a reasonable question is whether or not it performs similarly to (or better than) the proposed permutation-based procedure. Li et al. (1991) derived a method to combine test statistics after MI. Specifically, for m imputed data sets, two values are needed to compute the adjusted test statistic of interest: The mean of all observed test statistics from Pearson's chi-square test,

$$X^2_{avg} = \frac{\sum_{\ell=1}^{m} X^2_\ell}{m},$$

and an estimate of the increase in variance,

$$r = \left(1 + \frac{1}{m}\right) \frac{\sum_{\ell=1}^{m} \left(\sqrt{X^2_\ell} - \overline{\sqrt{\boldsymbol{X}^2}}\right)^2}{m-1},$$

where $\boldsymbol{X}^2 = (X_1^2, \ldots, X_m^2)$, the vector of m test statistics from each multiply-imputed data set. In other words, the second term of r is the sample variance of the square roots of the observed test statistics.

Then, the test statistic is given as

$$X^{2*} = \frac{\frac{X_{avg}^2}{d} - \frac{m+1}{m-1} r}{1+r},$$

where d is the degrees of freedom for a given test statistic from a multiply-imputed data set (here, $d = 1$).

Finally, the test statistic is F-distributed with d and $\nu = d^{\frac{-3}{m}}(m-1)\left(1+\frac{1}{r}\right)^2$ degrees of freedom:

$$p^* = P\left(F_{d,\nu} > X^{2*}\right)$$

In order to assess the asymptotic behavior of MI in this context, simulation is carried out in Section 2.3.4.

2.3 SIMULATION STUDIES

2.3.1 Defining the data structure

The following steps are used in algorithms *2-6* to define the data structure of interest. Specifically, they describe how to determine the distribution of (A, B, T) given a fixed sample size and distribution for (A, B), balanced treatment groups and a set amount of correlation between Y and T, expressed as an odds ratio (OR). This approach is taken since knowledge of (A, B, T) (as opposed to simply (Y, T)) is needed to assess how changing the distribution of (A, B) affects the simulation results. Note the final

distribution will be approximately distributed as (A, B, T) with approximately the fixed OR due to the requirement that cell counts be whole numbers.

Step 1: Determine the cell counts for the distribution of (Y, T) under fixed n and $(p_{11\cdot},\ p_{12\cdot},\ p_{21\cdot},\ p_{22\cdot})$, where $p_{ij\cdot} = P[(A, B) = (i, j)]$, balanced treatment groups and a fixed odds ratio between Y and T, $\frac{q_{11}q_{22}}{q_{12}q_{21}} \overset{set}{=} \gamma$, where $q_{kl} = P[(Y, T) = (k, l)]$:

(a) Given the conditions as in *Step 1*, find q_{11}, q_{12}, q_{21} and q_{22} analytically: First, solve for q_{22} in the quadratic equation $0.5\gamma p_{22\cdot} + (0.5 - 0.5\gamma - \gamma p_{22\cdot} - p_{11\cdot} - p_{12\cdot} - p_{21\cdot})q_{22} + (\gamma - 1)q_{22}^2 = 0$ under the restriction that $0 \leq q_{22} \leq 1$. Then, $q_{12} = 0.5 - q_{22}$ because of balanced treatment groups, $q_{21} = p_{22\cdot} - q_{22}$ because fixed (A, B) fixes the margins of Y, and $q_{11} = 0.5 - q_{21}$, again by balanced treatment groups.

(b) Allot $\lfloor nq_{22} \rfloor = m_1$ subjects to the cell corresponding to $Y = T = 2$, where $\lfloor\ \rfloor$ represents the floor function.

(c) Allot $\lfloor (q_{21} + q_{22})n \rfloor - m_1 = m_2$ subjects to the cell corresponding to $Y = 2$, $T = 1$.

(d) Allot $\frac{n}{2} - m_1$ subjects to the cell corresponding to $Y = 1$, $T = 2$.

(e) Allot remaining subjects to the cell corresponding to $Y = T = 1$.

Step 2: Based on the cell counts obtained in *Step 1* under a fixed OR between Y and T, derive the cell counts for the distribution of (A, B, T) while maintaining the distribution of (A, B) and balanced treatment groups:

(a) From steps *1*(c) and *1*(b), the cell counts for $(A = B = 2, T = 1)$ and $(A = B = T = 2)$ are m_2 and m_1, respectively, since $A = B = 2 \iff Y = 2$ as defined in (2.1). From this, p_{221} and p_{222} are known.

(b) To distribute the remainder of patients into the remaining cells of (A, B, T) while maintaining balanced treatment groups, let $\theta_1 = P(Y = T = 1) = 0.5 - p_{221}$ and $\theta_2 = P(Y = 1, T = 2) = 0.5 - p_{222}$, so that $\theta = \frac{\theta_1}{\theta_1+\theta_2}$ is the proportion of remaining cells corresponding to $T = 1$.

(c) Allot $\lfloor np_{11\cdot}\theta \rfloor = m_3$ subjects to the cell corresponding to $A = B = T = 1$.

(d) Allot $\lfloor np_{11\cdot} \rfloor - m_3$ subjects to $A = B = 1, T = 2$.

(e) Allot $\lfloor np_{12\cdot}\theta \rfloor = m_4$ subjects to the cell corresponding to $A = 1, B = 2, T = 1$.

(f) Allot $\lfloor np_{12\cdot} \rfloor - m_4$ subjects to the cell corresponding to $A = 1, B = T = 2$.

(g) Among the remaining $np_{21\cdot}$ subjects, assign $\frac{n}{2} - m_2 - m_3 - m_4$ to the cell corresponding to $A = 2, B = T = 1$ and the rest to the cell corresponding to $A = 2, B = 1, T = 2$.

2.3.2 Bias and variance of point estimates of the distribution of (Y, T) for all imputation procedures

The empirical bias and variance in the estimates of the cell probabilities of (Y, T) due to each imputation procedure can be quantified by utilizing *Algorithm 2*.

Algorithm 2: Calculate the empirical bias and variation of parameter estimates of the joint distribution of Y and T for all imputation procedures outlined in Section 2.2.3 under varying levels of correlation between Y and T (expressed as an odds ratio).

Step 1: For each sample, $X^{(d)} \sim MULTI(\mathbf{p})$; $d = 1, \ldots, D$, of size n established in Section 2.3.1:

(a) Make a fixed percentage of the data MCAR.

(b) Impute missing values using one of the procedures outlined in Section 2.2.3 and calculate Y appropriately in each instance.

(c) Store $\hat{\mathbf{q}}^{(d)}$, the vector of observed probabilities of (Y, T) after imputation.

(d) Calculate the vector of biases of $\hat{\mathbf{q}}^{(d)}$ for the d^{th} data set as $\boldsymbol{\delta}^{(d)} = \left(\left[\hat{q}_{11}^{(d)} - q_{11} \right], \left[\hat{q}_{12}^{(d)} - q_{12} \right], \left[\hat{q}_{21}^{(d)} - q_{21} \right], \left[\hat{q}_{22}^{(d)} - q_{22} \right] \right)$.

Step 2: Calculate the vector of average biases from estimating (Y, T) after imputation as $\frac{1}{D} \sum_{d=1}^{D} \boldsymbol{\delta}^{(d)}$.

Step 3: Calculate the vector of empirical standard deviations from estimating (Y, T) after imputation as $\sqrt{\frac{\sum_{d=1}^{D} \left[\hat{\mathbf{q}}^{(d)} - \bar{\hat{\mathbf{q}}} \right]^2}{D-1}}$, where $\bar{\hat{\mathbf{q}}}$ is the vector of means of $\hat{\mathbf{q}}^{(d)}$.

Here, the simulation sampled $D = 5000$ times from a set distribution with $n = 5000$ for each sample. When the OR between Y and T was 1, all imputation procedures had negligible bias (when comparing the empirical bias to the empirical SD; Table 2.1, bold) except those based on A and $A \,|\, T$, as these procedures are not valid even when $Y \perp T$, as shown in Sections 2.2.3.2 and 2.2.3.4. For ORs not equal to 1, only $A \,|\, (B, T)$ and $Y \,|\, T$ remained unbiased, as expected again due to the results in Section 2.2.3. Amongst the two aforementioned valid procedures, for a given cell of (Y, T), $A \,|\, (B, T)$ always had lower standard deviation (bold and red) compared to $Y \,|\, T$, making it more efficient (discussed in more detail in Section 2.3.3).

Table 2.1: Average empirical bias (*true* − *estimated*) and standard deviation (in parentheses) of the parameters of the joint distribution of Y and T after imputation. Cases with negligible bias are in bold. Invalid (biased) procedures are only included for the first parameter settings, as all other results are analogous. Results in red indicate lower SD when comparing the two valid imputation procedures.

$(p_{11\cdot}, p_{12\cdot}, p_{21\cdot}, p_{22\cdot})$[b]	Percent MCAR	Imputation procedure	Odds ratio[a]		
			0.25	1	2
(0.69, 0.11, 0.07, 0.13)	20	Y	-7.24e-03, 7.22e-03, 7.08e-03, -7.06e-03 (7.22e-03, 7.39e-03, 4.53e-03, 2.86e-03)	**1.88e-04, -5.94e-05, -6.39e-05, -6.43e-05** (7.22e-03, 7.16e-03, 3.82e-03, 3.79e-03)	3.65e-03, -3.74e-03, -3.79e-03, 3.87e-03 (7.20e-03, 7.19e-03, 3.38e-03, 4.31e-03)
		A	-1.42e-02, -2.20e-03, 1.40e-02, 2.36e-03 (7.11e-03, 7.26e-03, 4.05e-03, 2.30e-03)	-8.02e-03, -8.27e-03, 8.15e-03, 8.14e-03 (7.11e-03, 7.04e-03, 3.33e-03, 3.32e-03)	-5.19e-03, -1.13e-02, 5.06e-03, 1.14e-02 (7.06e-03, 7.05e-03, 2.83e-03, 3.80e-03)
		$A \mid B$	-3.90e-03, 3.86e-03, 3.74e-03, -3.70e-03 (7.11e-03, 7.28e-03, 4.48e-03, 2.63e-03)	**1.62e-04, -6.68e-05, -3.81e-05, -5.69e-05** (7.16e-03, 7.11e-03, 3.63e-03, 3.69e-03)	1.86e-03, -1.93e-03, -1.99e-03, 2.06e-03 (7.13e-03, 7.15e-03, 3.17e-03, 4.18e-03)
		$A \mid T$	-1.22e-02, -3.37e-03, 1.21e-02, 3.53e-03 (7.19e-03, 7.26e-03, 4.21e-03, 2.30e-03)	-8.02e-03, -8.27e-03, 8.14e-03, 8.15e-03 (7.14e-03, 7.08e-03, 3.41e-03, 3.39e-03)	-5.91e-03, -1.03e-02, 5.78e-03, 1.05e-02 (7.08e-03, 7.11e-03, 2.88e-03, 3.90e-03)
		$Y \mid T$[c]	**-8.98e-05, 1.17e-04, -6.82e-05, 4.13e-05** (7.54e-03, 7.40e-03, 5.00e-03, 2.84e-03)	**1.81e-04, -6.86e-05, -5.73e-05, -5.51e-05** (7.40e-03, 7.34e-03, 4.14e-03, 4.09e-03)	**-1.55e-04, 5.52e-05, 2.13e-05, 7.88e-05** (7.29e-03, 7.46e-03, 3.53e-03, 4.72e-03)
		$A \mid (B,T)$[c]	**-1.28e-04, 1.32e-04, -3.00e-05, 2.60e-05** **(7.19e-03, 7.35e-03, 4.60e-03, 2.72e-03)**	**1.52e-04, -6.16e-05, -2.82e-05, -6.21e-05** **(7.26e-03, 7.20e-03, 3.81e-03, 3.86e-03)**	**-1.24e-04, 6.24e-05, -9.48e-06, 7.16e-05** **(7.17e-03, 7.23e-03, 3.29e-03, 4.34e-03)**
(0.16, 0.29, 0.14, 0.41)	20	$Y \mid T$[c]	**2.85e-05, 4.67e-05, -1.32e-04, 5.72e-05** (6.59e-03, 7.54e-03, 7.23e-03, 5.53e-03)	**-5.24e-05, 1.27e-04, -2.34e-05, -5.10e-05** (7.40e-03, 7.37e-03, 6.67e-03, 6.70e-03)	**-7.08e-05, 4.77e-05, -2.27e-05, 4.58e-05** (7.32e-03, 7.05e-03, 6.13e-03, 6.93e-03)
		$A \mid (B,T)$[c]	**1.27e-05, 5.63e-05, -1.17e-04, 4.76e-05** **(6.30e-03, 7.42e-03, 6.96e-03, 5.36e-03)**	**-1.37e-04, 9.35e-05, 6.08e-05, -1.77e-05** **(7.17e-03, 7.07e-03, 6.35e-03, 6.45e-03)**	**-1.10e-04, 7.36e-05, 1.69e-05, 2.00e-05** **(7.19e-03, 6.74e-03, 5.87e-03, 6.65e-03)**
(0.69, 0.11, 0.07, 0.13)	70	$Y \mid T$[c]	**-3.70e-05, 8.21e-05, 7.63e-05, -1.21e-04** (9.86e-03, 8.05e-03, 8.18e-03, 4.72e-03)	**-2.87e-05, -2.72e-04, 2.35e-04, 6.53e-05** (9.13e-03, 9.15e-03, 6.78e-03, 6.68e-03)	**-1.31e-05, -7.32e-05, 1.58e-04, -7.13e-05** (8.89e-03, 9.62e-03, 5.85e-03, 7.64e-03)
		$A \mid (B,T)$[c]	**3.24e-05, 6.16e-05, 6.96e-06, -1.01e-04** **(8.19e-03, 7.71e-03, 6.11e-03, 4.13e-03)**	**1.70e-04, -2.19e-04, 3.67e-05, 1.26e-05** **(8.24e-03, 8.25e-03, 5.45e-03, 5.45e-03)**	**1.05e-04, -8.02e-05, 3.90e-05, -6.43e-05** **(8.30e-03, 8.24e-03, 4.96e-03, 5.90e-03)**
(0.16, 0.29, 0.14, 0.41)	70	$Y \mid T$[c]	**2.63e-05, 1.31e-05, -9.03e-05, 5.09e-05** (1.06e-02, 1.02e-02, 1.08e-02, 8.85e-03)	**-7.16e-05, 1.60e-07, 2.16e-05, 4.98e-05** (1.06e-02, 1.08e-02, 1.03e-02, 1.04e-02)	**-1.63e-04, -5.88e-05, 2.62e-04, -4.10e-05** (1.05e-02, 1.08e-02, 9.82e-03, 1.07e-02)
		$A \mid (B,T)$[c]	**1.32e-04, 4.64e-05, -1.96e-04, 1.76e-05** **(8.94e-03, 9.69e-03, 9.31e-03, 8.24e-03)**	**-1.34e-04, 2.88e-05, 8.39e-05, 2.12e-05** **(9.65e-03, 9.61e-03, 9.24e-03, 9.21e-03)**	**-3.50e-05, -1.96e-06, 1.35e-04, -9.78e-05** **(9.64e-03, 9.41e-03, 8.90e-03, 9.39e-03)**

[a] OR for Y across levels of T

[b] Corresponds to p_{ijl}, where A corresponds to i, B to j, T to l

[c] Valid imputation procedure

2.3.3 Inflated and corrected type-I error rates for Pearson's chi-square test on a singly-imputed data set

With regard to type-I error, naïve use of Pearson's test after imputation underestimates the variation in the data, thus increasing the rate of rejection beyond the expected level of α. This finding stems from the fact that the area under the tail of the χ^2 distribution increases with increasing variance, so there is more area to the right of a given critical value compared to the distribution when data are fully-observed.

In the remainder of the paper, two versions of the adjusted critical values are utilized: First, for each sample, the empirical distribution of the test statistic under the null is used to find the corrected critical value specific to that sample. Second, for a large number of samples from a set distribution for (A, B), the *average* corrected critical value is calculated and subsequently used as the critical value for any sample from that distribution. The motivation is to illustrate that the inflation in variation due to imputation depends only on the amount of missingness and the distribution of (A, B). Specifically, showing Pearson's test attains the nominal type-I error rate when using the average adjusted critical value specific to (A, B) for *any* sample from that distribution indicates other factors are not affecting the variation in the data.

Algorithm 3 is used to estimate the uncorrected rate of rejection under $Y \,|\, T$ and $A \,|\, (B, T)$, with these results reported in columns 3 and 5 of Table 2.2. In addition to the empirical standard deviation previously reported, these uncorrected rates confirm the intuitive notion that imputation using $A \,|\, (B, T)$ is more efficient than $Y \,|\, T$ in that the type-I error after imputing with $A \,|\, (B, T)$ was always smaller than the respective value after $Y \,|\, T$.

Algorithm 3: Quantify the inflated type-I error rate in Pearson's chi-square test when imputation is ignored:

Step 1: For each sample, $X^{(d)} \sim MULTI(\mathbf{p})$; $d = 1, \ldots, D$, of size n under the null of $(A, B) \perp T$ and with a set percentage of data MCAR:

(a) Impute with a valid procedure, $Y \,|\, T$ or $A \,|\, (B, T)$ (see Sections 2.2.3.5 and 2.2.3.6), calculating Y appropriately in each case.

(b) Conduct Pearson's test for independence between Y and T and let $\lambda_1^{(d)} = I\left\{S^{(d)} \geq S_{1-\alpha}\right\}$, where $S^{(d)}$ is the observed test statistic and $S_{1-\alpha}$ is the naïve critical value associated with α if data the were completely observed.

Step 2: Calculate the inflated rate of rejection as $\frac{1}{D}\sum_{d=1}^{D} \lambda_1^{(d)}$.

Recalling *Algorithm 1* outlines the permutation-based correction to Pearson's test after imputation under $A \,|\, (B, T)$, the asymptotic type-I rate of this procedure is simulated using *Algorithm 4*. The algorithm samples many times from a fixed multinomial distribution, carries out *Algorithm 1* in each instance and notes whether or not the test rejects. The average of this binary measure then gives an estimate of α.

Algorithm 4: Determine the adjusted critical value for Pearson's chi-square test after imputation under $A \,|\, (B, T)$ based on the empirical distribution of the test statistic and show the use of this value results in the nominal type-I error rate, α, asymptotically:

Step 1: For each sample, $X^{(d)} \sim MULTI(\mathbf{p})$; $d = 1, ..., D$, of size n under the null of $(A, B) \perp T$ and with a set percentage of data MCAR:

(a) Impute with $A \,|\, (B, T)$ as outlined in Section 2.2.3.6 and calculate Y.

(b) Conduct Pearson's test for independence between Y and T and record $S^{(d)}$, the observed test statistic.

(c) For $g = 1, ..., G$:

(i) Permute T randomly to simulate the null of $(A, B) \perp T$.

(ii) Impute missing data with $A \,|\, (B, T)$ and calculate Y appropriately.

(iii) Conduct Pearson's test between Y and T and record $S_g^{(d)}$, the observed test statistic from the permuted and imputed data.

(d) Organize the set of G test statistics from *Step 1* (c), $\left\{S_1^{(d)}, S_2^{(d)}, \ldots, S_G^{(d)}\right\}$, in ascending order and find $S_{1-\alpha}^{(d)}$, the $(1-\alpha)^{\text{th}}$ percentile of the empirical distribution of the test statistic, which is the corrected critical value of interest.

(e) Let $\lambda_2^{(d)} = I\left\{S^{(d)} \geq S_{1-\alpha}^{(d)}\right\}$.

Step 2: Calculate the corrected rate of rejection as $\frac{1}{D}\sum_{d=1}^{D} \lambda_2^{(d)}$.

Lastly, since it is believed the corrected critical value depends only on the distribution of the data and percent of missingness, *Algorithm 5* finds the average corrected value over many samples from a set distribution and uses this value to show the test rejects at the expected rate of α in the long-run. In other words, for a given distribution and percent of missing data, there exists one true correction factor, which could hypothetically be known. Each corrected value obtained from a sample is an estimate of this true value.

Algorithm 5: Show the adjusted critical value of Pearson's chi-square test depends only on the distribution from which the sample is drawn and the percent of missing

data by estimating the average adjusted value and showing the asymptotic rate of rejection based on this estimate approaches α.

Step 1: For each sample, $X^{(d)} \sim MULTI(\mathbf{p})$; $d = 1, ..., D$, of size n under the null of $(A, B) \perp T$ and with a set percentage of data MCAR, conduct steps *1*(a)-(d) from *Algorithm 4* and store $S^{(d)}_{1-\alpha}$ in each instance.

Step 2: Find the average of all $S^{(d)}_{1-\alpha}$ as $S^{*}_{1-\alpha} = \frac{1}{D}\sum_{d=1}^{D} S^{(d)}_{1-\alpha}$.

Step 3: For each of $g = 1, ..., G$ samples from the same distribution as defined in *Step 1*:

(a) Impute using $A \,|\, (B, T)$ as in Section 2.2.3.6 and calculate Y.

(b) Conduct Pearson's test between Y and T and let $\lambda_3^{(g)} = I\left\{S^{(g)} \geq S^{*}_{1-\alpha}\right\}$, where $S^{(g)}$ is the observed test statistic.

Step 4: Calculate the corrected rate of rejection based on the average adjusted critical value as $\frac{1}{G}\sum_{g=1}^{G} \lambda_3^{(g)}$.

The results from algorithms *3*, *4* and *5* are given in Table 2.2, with $D = 2000$ and $n = 1000$. $G = 3000$ for each of the D samples in *Algorithm 4*, and for *Step 1* in *Algorithm 5* (which references *Algorithm 4*), $G = 3000$ (for determining the empirical distribution of the test statistic), while for *Step 3*, $G = 2000$ (number of samples from the distribution).

Looking at the last column of Table 2.2, the higher efficiency of $A \,|\, (B, T)$ is again illustrated in that the average corrected critical value for $A \,|\, (B, T)$ was always smaller than that for $Y \,|\, T$ (column 4). Specifically, the more the variation is increased due to imputation, the larger the critical value needs to be to ensure the test rejects at the nominal level. Thus, smaller adjusted critical values imply a more efficient imputation procedure. This increased efficiency under $A \,|\, (B, T)$ comes from the fact

Table 2.2: Uncorrected and corrected type-I error rates for Pearson's chi-square test of independence between Y and T after valid imputation (nominal type-I error rate is $\alpha = 0.05$).

		$Y \mid T$		$A \mid (B, T)$		
$(p_{11\cdot}, p_{12\cdot}, p_{21\cdot}, p_{22\cdot})$[a]	Percent MCAR	Uncorrected	Corrected[b] (Corrected critical value)	Uncorrected	Corrected I[c]	Corrected II[d] (Average corrected critical value)
(0.69, 0.11, 0.07, 0.13)	20	0.0980	0.0470 (5.568)	0.0875	0.0560	0.0525 (4.748)
(0.44, 0.05, 0.42, 0.09)	20	0.1165	0.0575 (5.568)	0.0650	0.0485	0.0520 (4.512)
(0.16, 0.29, 0.14, 0.41)	20	0.1035	0.0510 (5.568)	0.0955	0.0545	0.0455 (5.052)
(0.26, 0.22, 0.23, 0.29)	20	0.0920	0.0435 (5.568)	0.0780	0.0505	0.0500 (4.891)
(0.69, 0.11, 0.07, 0.13)	70	0.3320	0.0475 (15.488)	0.2290	0.0425	0.0560 (10.165)
(0.44, 0.05, 0.42, 0.09)	70	0.3230	0.0460 (15.488)	0.1920	0.0495	0.0485 (8.643)
(0.16, 0.29, 0.14, 0.41)	70	0.3300	0.0515 (15.488)	0.2720	0.0510	0.0575 (12.041)
(0.26, 0.22, 0.23, 0.29)	70	0.3415	0.0510 (15.488)	0.2385	0.0490	0.0495 (10.982)

[a] Corresponds to p_{ijl}, where A corresponds to i, B to j, T to l

[b] Using closed-form adjusted critical value established by Wang (2006)

[c] Adjusted critical value determined for each of $d = 1, \ldots, D$ data sets (sampled from the same population)

[d] Adjusted rate of rejection for D data sets based on average corrected critical value from D data sets sampled from the same distribution

that Y is formed as a function of A and B because of contextual information suggesting A and B are correlated. From this, more information is retained if A is imputed based on knowledge of both B and T, as opposed to first calculating Y and imputing this variable based only on T.

Additionally, assuming Y is missing when A is missing and imputing all missing Y values discards instances where Y is already determined based on the knowledge of B. Specifically, given definition (2.1), knowledge that $B = 1$ indicates $Y = 1$ even if A is missing. Thus, ignoring this information unnecessarily imputes known Y values. In contrast, imputation under $A \,|\, (B, T)$ retains this information: Even if A was imputed, Y is calculated after the fact, so that despite the value of A, if $B = 1$ then $Y = 1$. Thus, the only observations that are "recognized" as imputed in the final data set are those for which A was missing and $B = 2$. Because of this, a smaller percentage of observations is acknowledged as imputed under $A \,|\, (B, T)$ compared to $Y \,|\, T$, making the former more efficient.

The above notions imply that for $A \,|\, (B, T)$, the rate of rejection will depend not only on the percentage of missing data, but also on the true cell probabilities that define the joint multinomial distribution of A and B. This concept is directly related to the definition of Y as well. For example, when Y is defined as in (2.1), if $B = 1$ and A is missing, then how A is imputed is not important – Y will be 1 regardless. The opposite would be true if Y were defined as, for example,

$$Y = \begin{cases} 2 & \text{if } A = B = 1 \\ 1 & \text{o.w.} \end{cases} . \tag{2.11}$$

Now the distribution of (A, B) impacts the rate of rejection in a different way: The cells with $B = 1$ dictate the amount of imputation realized in the data set.

Somewhat obviously, neither the distribution of (A, B) nor the definition of Y

affect the amount of inflation after imputation under $Y \mid T$. This is again because all values of Y are set to missing if A is missing, so no information is maintained based on the distribution of (A, B)/definition of Y. This notion is confirmed by observing that for imputation under $Y \mid T$, the simulated inflation in the rate of rejection depended only on the percent of missing data (Table 2.2, column 3). Specifically, for 20% of data MCAR, the rate was approximately 0.10 despite the distribution of (A, B), and increased to about 0.33 for 70% missing data.

In contrast, the inflation in rejection after imputation via $A \mid (B, T)$ depended not only on the percent of missingness, but also on the distribution of (A, B). For example, with 70% of data MCAR and cell probabilities for (A, B) of $(p_{11\cdot}, p_{12\cdot}, p_{21\cdot}, p_{22\cdot}) = (0.44, 0.05, 0.42, 0.09)$, the rate was 0.192, but this increased to 0.272 when the distribution changed to (0.16, 0.29, 0.14, 0.41). This finding is in the expected direction since $P(B = 2)$ increased from 0.14 to 0.7, and given the way Y is defined, the cells with $B = 2$ will be those recognized as imputed if A is missing. This pattern is illustrated in detail in Table 2.3. For the marginal probability that $B = 2$, there is a monotonically-increasing pattern with the uncorrected type-I error, while for $A = 2$, there is no pattern.

Lastly, the results of *Algorithm 4* confirmed Pearson's test rejects at the expected α-level asymptotically when the empirically-corrected critical value is used, as inflated rates of rejection were reduced to approximately 0.05 in all cases (Table 2.2, column 6). Additionally, the simulation based on *Algorithm 5* (which considers the *average* corrected critical value) showed there is some hypothetical corrected critical value that depends only on the distribution of the data and percent of data MCAR: When the average corrected critical value was used instead of the value specific to each data set, the rates of rejection were still at the nominal level of 0.05 (last column of Table 2.2).

Table 2.3: Comparison of the uncorrected type-I error rate of Pearson's test based on the marginal probabilities of the components of the complex outcome.

$(p_{11\cdot}, p_{12\cdot}, p_{21\cdot}, p_{22\cdot})$	$P(A=2)$	% MCAR	Uncorrected α
(0.69, 0.11, 0.07, 0.13)	0.20	70	0.2290
(0.44, 0.05, 0.42, 0.09)	0.51	70	0.1920
(0.26, 0.22, 0.23, 0.29)	0.52	70	0.2385
(0.16, 0.29, 0.14, 0.41)	0.55	70	0.2720

$(p_{11\cdot}, p_{12\cdot}, p_{21\cdot}, p_{22\cdot})$	$P(B=2)$	% MCAR	Uncorrected α
(0.44, 0.05, 0.42, 0.09)	0.14	70	0.1920
(0.69, 0.11, 0.07, 0.13)	0.24	70	0.2290
(0.26, 0.22, 0.23, 0.29)	0.51	70	0.2385
(0.16, 0.29, 0.14, 0.41)	0.70	70	0.2720

2.3.4 Comparison of permutation-based method to multiple imputation

As outlined previously, the relative performance of multiple imputation to the proposed permutation-based method for correcting Pearson's test after single imputation is of interest. As such, the inferential MI procedure described in Section 2.2.5.3 was carried out for a large number of samples from a given distribution ($D = 2000$, with $n = 1000$ for each sample) under $A \,|\, (B, T)$. Whether or not the adjusted test rejected was recorded and the overall rate was calculated over these 2000 samples. Similar to Table 2.2, Table 2.4 shows the uncorrected and corrected α-levels across various parameter settings. This analysis was conducted for $m = 3$ and $m = 10$ to assess the sensitivity of the results (if any) to the number of imputed data sets. These choices were based on Li et al. (1991), who used $m = 2, 3, 5$ and 10. For the uncorrected rate, one instance of imputation was conducted (for each of the D samples), and a critical value of 3.84 was used for Pearson's test.

Table 2.4 shows that for 20% missing data, MI with both $m = 3$ and 10 performed as expected − the nominal rejection rate (0.05) was attained in both cases, although

Table 2.4: Uncorrected and corrected type-I error rates for Pearson's test between Y and T after multiple imputation under $A \,|\, (B, T)$ (nominal type-I error rate is $\alpha = 0.05$).

		$m = 3$		$m = 10$	
$(p_{11\cdot}, p_{12\cdot}, p_{21\cdot}, p_{22\cdot})$[a]	Percent MCAR	Uncorrected	Corrected	Uncorrected	Corrected
(0.69, 0.11, 0.07, 0.13)	20	0.0815	0.0500	0.0785	0.0480
(0.44, 0.05, 0.42, 0.09)	20	0.0695	0.0430	0.0695	0.0525
(0.16, 0.29, 0.14, 0.41)	20	0.0785	0.0400	0.0880	0.0500
(0.26, 0.22, 0.23, 0.29)	20	0.0805	0.0500	0.0940	0.0515
(0.69, 0.11, 0.07, 0.13)	70	0.2375	**0.1140**	0.2280	**0.1305**
(0.44, 0.05, 0.42, 0.09)	70	0.1825	**0.0970**	0.1765	**0.1010**
(0.16, 0.29, 0.14, 0.41)	70	0.2840	**0.1340**	0.2675	**0.1310**
(0.26, 0.22, 0.23, 0.29)	70	0.2485	**0.1095**	0.2390	**0.1180**

[a] Corresponds to p_{ijl}, where A corresponds to i, B to j, T to l

there was potential downward bias for some settings when $m = 3$. However, for 70% missingness, the performance was not optimal despite the value of m (values in bold). For $m = 3$, uncorrected rates ranging from about 0.18 to 0.28 were corrected to 0.097 and 0.134, respectively. Similarly, for $m = 10$, uncorrected rates from approximately 0.18 to 0.27 were corrected to 0.10 and 0.13. As such, this method provides biased results with large rates of missingness.

These conclusions are consistent with those reached by the authors, who observed both downward and upward bias in the type-I error rate depending on the chosen α-level, amount of missing data and number of imputed data sets (Li et al., 1991). Their simulations did not consider instances with more than 50% missing data, as they claimed in practice it is unlikely to observe higher levels than this. However, even for $\leq 50\%$ missing data, they did detect bias, as noted above.

In conclusion, there is no motivation to use MI as presented by Li et al. (1991)

in this instance, as it behaves erratically while the permutation-based method gives reliable results despite the structure of the data and amount of missingness.

2.3.5 Power of valid imputation procedures

In addition to efficiency, valid imputation procedures may also be compared with regard to their ability to reject the null when they should:

Algorithm 6: Estimate the power of Pearson's chi-square test when a) using the closed-form corrected critical value after imputation under $Y \,|\, T$ (Wang, 2006), b) using the the empirically-adjusted critical value after imputation under $A \,|\, (B, T)$ (i.e., by using the method in *Algorithm 1*) or c) using the average empirically-adjusted critical value after imputation under $A \,|\, (B, T)$ (i.e., by using the method in *Algorithm 5*):

Step 1: Generate a distribution for (A, B, T) as in Section 2.3.1, so that Y and T are correlated according to a fixed odds ratio, γ.

Step 2: For each sample, $X^{(d)}$; $d = 1, ..., D$, of size n from the distribution of (A, B, T) established in *Step 1*:

(a) Make a fixed percentage of the data MCAR.

(b) Impute with either $Y \,|\, T$ or $A \,|\, (B, T)$ (see Sections 2.2.3.5 and 2.2.3.6) and calculate Y as appropriate.

(c) Conduct Pearson's test for independence between Y and T and record $S^{(d)}$, the observed test statistic.

(d) If imputation under $Y \,|\, T$ was used: Let $\lambda_4^{(d)} = I\left\{\frac{S^{(d)}}{\pi_c^{-1}+1-\pi_c} \geq S_{1-\alpha}\right\}$, where $S_{1-\alpha}$ is the naïve critical value associated with α if data were

completely observed. $\frac{S^{(d)}}{\pi_c^{-1}+1-\pi_c}$ is the closed-form adjusted test statistic as derived by Wang (2006). Else, go to *Step 2*(e).

(e) If imputation under $A \,|\, (B,T)$ was used, conduct steps *1*(c)-(d) of *Algorithm 4*, store the value of $S^{(d)}_{1-\alpha}$ and let $\lambda_4^{(d)} = I\left\{S^{(d)} \geq S^{(d)}_{1-\alpha}\right\}$.

Step 3: If imputation under $A \,|\, (B,T)$ was used, conduct *Step 2* of *Algorithm 5* based on the stored values from the previous step and calculate $\lambda_5^{(d)} = I\left\{S^{(d)} \geq S^{*}_{1-\alpha}\right\} \forall\ d$.

Step 4: Calculate the applicable power values as $\frac{1}{D}\sum_{d=1}^{D}\lambda_4^{(d)}$ and $\frac{1}{D}\sum_{d=1}^{D}\lambda_5^{(d)}$.

Based on $D = 2000$ samples ($n = 5000$), for all values of the OR between Y and T (except 1), the power was higher for imputation under $A \,|\, (B,T)$ than under $Y \,|\, T$ when using either the empirically-adjusted critical value specific to each data set or the average across the 2000 data sets (Table 2.5). In some instances, the increase in power was large − up to 17 percentage points higher. When the OR was 1, all procedures rejected around 5% of the time, as expected.

Table 2.5: Power (percent) of Pearson's test after imputation using valid procedures. Largest values within a given set of parameters and odds ratio column are denoted in bold.

$(p_{11\cdot}, p_{12\cdot}, p_{21\cdot}, p_{22\cdot})$[b]	Percent MCAR	Imputation procedure/ Correction factor	Odds ratio[a] 0.5	0.75	1	1.25	1.5	1.75	2
(0.69, 0.11, 0.07, 0.13)	20	$Y \mid T$[c]	100	81.75	5.55	59.75	97.50	99.90	100
		$A \mid (B,T)$[d]	100	**86.85**	5.60	63.10	**99.15**	**100**	100
		$A \mid (B,T)$ (average)[e]	100	**86.85**	5.30	**64.85**	98.90	99.95	100
(0.16, 0.29, 0.14, 0.41)	20	$Y \mid T$[c]	100	98.95	5.60	87.75	100	100	100
		$A \mid (B,T)$[d]	100	**99.25**	4.85	92.05	100	100	100
		$A \mid (B,T)$ (average)[e]	100	99.20	4.25	**92.85**	100	100	100
(0.69, 0.11, 0.07, 0.13)	70	$Y \mid T$[c]	98.20	40.90	5.15	25.40	66.20	89.15	97.70
		$A \mid (B,T)$[d]	99.65	**58.55**	5.60	**36.25**	**83.35**	98.25	**99.85**
		$A \mid (B,T)$ (average)[e]	**99.99**	55.20	4.45	35.75	83.10	**98.50**	**99.85**
(0.16, 0.29, 0.14, 0.41)	70	$Y \mid T$[c]	100	70.10	5.65	47.75	93.90	99.55	99.95
		$A \mid (B,T)$[d]	100	**81.05**	4.80	**59.25**	**97.95**	**100**	100
		$A \mid (B,T)$ (average)[e]	100	80.35	5.25	58.15	97.80	**100**	100

[a] OR for Y across levels of T
[b] Corresponds to p_{ijl}, where A corresponds to i, B to j, T to l
[c] Using closed-form adjusted critical value established by Wang (2006)
[d] Using adjusted critical value specific to each of $d = 1, \ldots, D$ data sets (sampled from the same population)
[e] Using average adjusted critical value from D data sets sampled from the same distribution

2.4 APPLICATION IN A BREAST CANCER CLINICAL TRIAL

The B-41 protocol of the National Surgical Adjuvant Breast and Bowel Project was a randomized, phase-3 clinical trial studying the efficacy of trastuzumab plus lapatinib compared to trastuzumab alone in neoadjuvant therapy for breast cancer (Robidoux et al., 2013). Additional detail was given previously in the introduction. As an illustration of the proposed method, a subset of the outcomes of this study − namely, the results of the nodal dissection − was considered.

After patients received chemotherapy to shrink a primary tumor before surgery, they underwent one of three procedures based on physician preference: 1) an *axillary dissection*, where a sample of SN and AN were removed from the arm closest to the tumor, but the types of nodes were not distinguished, 2) the removal of all SN and AN, where SN were identified by a tracer/dye, or 3) initial detection and removal of only SN by use of a tracer/dye. If any SN were positive, AN were removed as well. If instead they were negative, no further removal of nodes was undertaken. Here, the complex outcome, Y, represents whether a patient had either no cancer in any lymph nodes ($Y = 2$) or at least one node with cancer ($Y = 1$).

Women in case (2) are known as "completers" and are used to impute the missing AN status for those in group (3). For those in case (1), since a given sample includes both SN (B) and AN (A), if all nodes are negative, let $A = B = 2$. Conversely, if at least one is positive, $A = B = 1$. Given $Y = 1$ when any node has cancer, this assumption that $A = B = 1$ correctly classifies a woman according to the definition of Y.

Up to this point, data were assumed MCAR. Here, since missing AN status depends on the SN biopsy, data are instead missing at random. However, because imputation of AN is done conditionally on SN, this relationship is accounted for

Table 2.6: Classification of neoadjuvant breast cancer data ($n = 331$) by treatment and sentinel and axillary node statuses (no nodes or at least one node positive for cancer). Missing data is denoted by $\cdot$.

		Trastuzumab alone				Trastuzumab + lapatinib			
		AN				AN			
		0	> 0	$\cdot$	Total	0	> 0	$\cdot$	Total
SN	0	60	2	69	131	80	1	55	136
	> 0	9	30	1	40	6	17	1	24
	Total	69	32	70	171	86	18	56	160

and point estimates are unbiased. Additionally, this structure does not affect the proposed method of permuting the treatment vector under the null, as missingness does not depend on treatment (since subjects were randomized). As described in Section 2.1, this simplified form of the data assumes no other predictors determine the missingness of A. However, if $\mathbf{Z}$ represents a set of covariates believed to affect the probability of a missing observation, imputation would be conducted conditionally on $\mathbf{Z}$ (i.e., as $A \,|\, (B, T, \mathbf{Z})$).

Table 2.6 classifies study participants by treatment and nodal status (missing data is denoted by $\cdot$). Sixty-nine women (20.8%) had both their SN and AN removed and biopsied, and estimates based on these women were used to impute those with missing AN ($n = 126$ (38.1%)). Additionally, 136 (41.1%) women had axillary dissections that were a sample of both SN and AN, with both types of nodes assumed to have the same status.

After imputation and without correction, the probability of no cancer in any lymph nodes given trastuzumab alone was 0.725, while that for trastuzumab and lapatinib combined was 0.819. Pearson's test for independence between Y and treatment resulted in a p-value of 0.043, indicating the addition of lapatinib significantly improved the rate of cancer-free lymph nodes over trastuzumab alone.

After simulating the distribution of the test statistic under the null, the corrected p-value was defined as the proportion of test statistics larger than observed, uncorrected test statistic. The p-value increased to 0.107, indicating there was no significant difference amongst treatments with regard to the rate of cancer found in the lymph nodes.

Again, this example was a simplified version of the analysis that would take place in clinical practice, where other covariates that could affect missingness would be considered. It serves to illustrate, however, the importance of correcting inference after imputation, as conclusions did change in this instance.

2.5 DISCUSSION

This chapter addressed the imputation of a complex outcome and the associated adjustment to Pearson's test for independence. In an attempt to build on Wang's (2006) finding for a simple binary outcome, it was determined the closed-form theoretical extension to this higher-dimensional problem may not be attainable. In light of this, a data-driven, permutation-based method that estimates the empirical distribution of the test statistic under the null was proposed. With simulation, this method's ability to provide correct inference based on an adjusted critical value was confirmed. It was also shown the imputation scheme of $A \,|\, (B, T)$ is more efficient and has greater power than the naïve method of $Y \,|\, T$. Lastly, a comparison of the suggested procedure to multiple imputation was undertaken, which demonstrated the superiority of the permutation-based method given its robustness to the percentage of missing data.

One weakness of these findings is the lack of a closed-form correction factor.

Ideally, the distribution of the test statistic could be deduced, such that Pearson's test could be adjusted without the use of the test statistic's empirical distribution, which would save some computational time. Additionally, these findings could be extended to the case where both A and B are subject to missingness (the results are likely analogous − imputation must always be done conditionally on all other variables in order to obtain consistent estimates) and/or where A, B and Y have more than two levels (again, the extension is likely trivial). Lastly, this paper only addressed the simplified case where Y was missing whenever A was missing, without utilizing the instances where the value of B determines Y, despite the imputed value of A.

Further work could explore this scenario where the y-values for the aforementioned subjects are treated as observed, which would affect the estimates used for imputing the remaining, "truly" missing values under $Y \,|\, T$. In this setting, the estimates used for imputation would be more precise as they would employ a larger sample size. However, only T is used for stratification, contrasting $A \,|\, (B, T)$, which stratifies on both B and T. It is possible, then, that there is a trade-off in the precision gained by the increased sample size in the $Y \,|\, T$ case versus the gain in precision due to more specific stratification under $A \,|\, (B, T)$. As such, it is likely the solution to this problem depends on a number of factors, including the distribution of A and B, the definition of Y and the percent of missingness − namely, the percent of subjects for which $A = \cdot$ and $B = 1$, as this is the case where $Y = 1$ despite the imputed value of A, given definition (2.1). It is also not clear whether Wang's (2006) closed-form correction factor readily holds in this setting for $Y \,|\, T$.

Lastly, as presented in Section 2.4, when the missingness of A depended on B but not T, imputation under $A \,|\, (B, T)$ was still valid. How this type of missing at random structure affects imputation via $Y \,|\, T$ when values of Y are assumed known

given B (as discussed in the previous paragraph) is also of interest. Specifically, if there are instances where $Y \,|\, T$ is more efficient than $A \,|\, (B, T)$, yet bias would be introduced if data were actually MAR, this could indicate the "safer" choice in general is $A \,|\, (B, T)$.

The results presented here highlight the importance of differentiating between a simple and complex binary outcome when data are missing. Specifically, where imputation occurs (i.e., at the A/B or Y level) affects efficiency and the distribution of the test statistic. Thus, the use of Wang's (2006) result under $A \,|\, (B, T)$ leads to incorrect inference. It is notable that in this context, imputation under $A \,|\, (B, T)$ results in higher efficiency and power, and that the suggested permutation-based method for correcting type-I error outperforms multiple imputation.

3.0 A MODIFIED EM ALGORITHM FOR CONTINGENCY TABLE ANALYSIS WITH MISSING DATA

3.1 INTRODUCTION

Maximum likelihood (ML) is a parameter estimation method, popular due to its simple implementation and favorable properties (Pawitan, 2001). Specifically, estimates are consistent and efficient − they obtain the Cramér-Rao lower bound asymptotically. Further, ML estimates for data with missing values may be obtained using the expectation maximization (EM) algorithm (Dempster et al., 1977). In general, the maximum likelihood method requires that if data are not presumed at least MAR, the missing-data mechanism is modeled or assumed known. Bias in the parameter estimate of interest may occur when such assumptions are misspecified. Additional detail about this method has been given previously in Sections 1.3.4.1 and 1.3.4.2.

The *EM algorithm* was formally established in the context of missing data by Dempster et al. (1977). It provides an alternative optimization routine to the Newton-Raphson algorithm and its extensions, with the advantage of being a stable, iterative procedure where the likelihood function is always increasing. Under general conditions, it always converges to the global maximum (maximum likelihood estimate (MLE)), with the potential exception of convergence to a local maximum

if it exists. This last fact indicates the initial value is important in a multimodal likelihood (Wu, 1983; Little and Rubin, 2002).

The fact that the EM algorithm always converges is due to the ability to separate the observed-data log-likelihood, $\ell(\theta; y_{obs})$, into the difference of two terms, Q and H, where y_{obs} represents the set of observed data, and $\theta \in \Omega_\theta$ is the unknown parameter to be estimated (see Section 3.1.1). Jensen's inequality guarantees H decreases with each iteration, so that the focus of the algorithm is on maximizing Q at each step (Dempster et al., 1977). Specifically, when Q increases, so does $Q - H$, and therefore $\ell(\theta; y_{obs})$. In other words, given a convergence criterion, the algorithm will approach a mode of the likelihood. Although the number of iterations required for convergence is often larger, the EM algorithm is preferred over Newton-Raphson-type methods because of its stability.

Presented here is a modification of the general EM algorithm, followed by its application to contingency table analyses under varying model structures and with one variable subject to missingness. The assumption is that consistent initial estimates of the model parameters are attainable. The case where the data used to obtain these estimates (the "external" data) is additionally available is also considered.

Briefly, the modified algorithm combines information from the data set subject to missingness with the initial estimates from the external data − and possibly external data itself − to produce consistent, yet potentially more efficient, estimates than those from the external data alone (i.e., the initial estimates). This is true regardless of the missing data mechanism (i.e., even if data are missing not at random). For certain model structures, the initial estimates are sufficient to provide this increase in efficiency, given the data set subject to missingness is at least slightly larger than the data set that provided the initial estimates. Essentially, the algorithm combines the initial estimates/external data with the data that, on its own, would produce

biased estimates (hence the gain in efficiency), but without inducing bias in these final estimates.

The assumptions about consistent initial estimates and the availability of external data are feasible for study designs that purposefully allow missing data so as to save money and/or resources (details below). Alternatively, estimates might be obtained from a data set believed to follow the same distribution as the data under study, although the validity of this scenario may be difficult to verify.

There exist two types of studies where the design includes purposeful missingness while allowing consistent estimation of model parameters. In the first type, a random sub-sample is drawn under the assumption that any missing observations in that sub-sample may be recovered (through subsequent interviews, diagnostic testing that was originally omitted, etc.). As such, this sub-sample will be representative of the entire sample, and thus estimates based on it will be consistent. In the second scenario, the sample is initially randomly divided into two sections: In the first, it is assumed (by use of valuable incentives, necessary resources, etc.) that all observations are attainable. In the second (collected with less-valuable incentives, a reduced budget, etc.), data are allowed to be missing. In this case, the first division of the data provides consistent estimation. In each of these frameworks, the missing portion of the sample may be missing not at random.

The rest of this chapter is structured as follows: In the remainder of the introduction, an outline of the general EM algorithm is provided, including the cases where the missing data mechanism is ignorable or assumed known. In Section 3.2.1, the modified EM algorithm is introduced in the regression framework. A discussion of the general and modified EM within the context of contingency table analyses follows. Finally, application of the modified EM algorithm under three different discrete data model structures is considered. In Section 3.3, simulation studies are presented

to illustrate the findings of Section 3.2. Of interest is distinguishing between the model structures that only require consistent initial estimates in order to increase efficiency, and those that additionally require the external data itself. Finally, Section 3.4 applies the proposed method to a data set regarding the effect of radiation on the length of survival after surgery due to ovarian cancer (< 10 years vs. ≥ 10 years) after controlling for stage of cancer (low vs. high).

3.1.1 The general EM algorithm for missing data

For a single unknown parameter (similar concepts apply to multiple parameters), denote the likelihood function by $L(\theta; y) = L(\theta; y_{obs}, y_{mis})$, where y_{obs} are the observed values, y_{mis} the missing values, θ exists in the parameter space Ω_θ and $L(\theta; y) \propto f(y; \theta)$, the pdf of Y. The (natural) log-likelihood is denoted by $\ell(\theta; y) = ln[L(\theta; y)]$.

When data are subject to missingness, the observed data contain both the observed values, y_{obs}, and the missing data indicator, R, a random variable with pdf $f(r|y; \psi)$. The joint distribution of Y_{obs} and R is then used to determine the *full likelihood model*: For $\theta, \psi \in \mathbf{\Omega}_{\theta,\psi}$,

$$L_{full}(\theta, \psi; y_{obs}, r) \propto f(y_{obs}, r; \theta, \psi) = \int f(y_{obs}, y_{mis}; \theta) f(r \mid y_{obs}, y_{mis}; \psi) dy_{mis}$$

(Little and Rubin, 2002).

The EM algorithm provides a stable optimization procedure for maximizing the full likelihood when data are subject to missingness. Expressing $y_{i,\,obs}$ as $r_i y_i$ and $y_{i,\,mis}$ as $(1 - r_i)y_i$, the derivation that underlies the algorithm is as follows:

$$f(y, r; \theta, \psi) = f(y_{obs}, y_{mis}, r; \theta, \psi) = f(y_{obs}, r; \theta, \psi) f(y_{mis} \mid y_{obs}, r; \theta, \psi)$$

$$\Rightarrow f(y_{obs}, r; \theta, \psi) = \frac{f(y, r; \theta, \psi)}{f(y_{mis} \mid y_{obs}, r; \theta, \psi)}$$

$$\Rightarrow L(\theta, \psi; y_{obs}, r) = \prod_{i=1}^{n} \frac{f(y_i, r_i; \theta, \psi)}{f[(1-r_i)y_i \mid r_i y_i, r_i; \theta, \psi]}$$
$$= \prod_{i=1}^{n} \frac{f(y_i; \theta) f(r_i \mid y_i; \psi)}{f[(1-r_i)y_i \mid r_i y_i, r_i; \theta, \psi]}$$

$$\Rightarrow \ell(\theta, \psi; y_{obs}, r) = \sum_{i=1}^{n} ln\left\{ \frac{f(y_i; \theta) f(r_i \mid y_i; \psi)}{f[(1-r_i)y_i \mid r_i y_i, r_i; \theta, \psi]} \right\}$$
$$= \sum_{i=1}^{n} \Big\{ ln\Big[f(y_i; \theta) \Big] + ln\Big[f(r_i \mid y_i; \psi) \Big] -$$
$$ln\Big\{ f[(1-r_i)y_i \mid r_i y_i, r_i; \theta, \psi] \Big\} \Big\}. \tag{3.1}$$

Thus, using the EM algorithm, to maximize $\ell(\theta, \psi; y_{obs}, r)$, one may maximize the RHS of (3.1). Taking the conditional expectation of both sides of (3.1) with respect to the missing data given the observed data and the current estimates of θ and ψ, $\theta^{(t)}$ and $\psi^{(t)}$,

$$E\Big[\ell(\theta, \psi; y_{obs}, r) \mid y_{obs}, r; \theta^{(t)}, \psi^{(t)} \Big]$$
$$\equiv \ell(\theta, \psi; y_{obs}, r)$$
$$= \sum_{i=1}^{n} E\Big\{ ln\Big[f(y_i; \theta) \Big] + ln\Big[f(r_i \mid y_i; \psi) \Big] \mid r_i y_i, r_i; \theta^{(t)}, \psi^{(t)} \Big\} -$$
$$\sum_{i=1}^{n} E\Big\{ ln\Big\{ f[(1-r_i)y_i \mid r_i y_i, r_i; \theta, \psi] \Big\} \mid r_i y_i, r_i; \theta^{(t)}, \psi^{(t)} \Big\} \tag{3.2}$$
$$:= Q\left[\theta, \psi \mid \theta^{(t)}, \psi^{(t)} \right] - H\left[\theta, \psi \mid \theta^{(t)}, \psi^{(t)} \right],$$

where $Q\left[\theta, \psi \,|\, \theta^{(t)}, \psi^{(t)}\right] = \sum_{i=1}^{n} E\Big\{ln\Big[f(y_i;\theta)\Big] + ln\Big[f(r_i \,|\, y_i;\psi)\Big] \;|\; r_i y_i, r_i; \theta^{(t)}, \psi^{(t)}\Big\}$ and $H\left[\theta, \psi \,|\, \theta^{(t)}, \psi^{(t)}\right] = \sum_{i=1}^{n} E\Big\{ln\Big\{f[(1-r_i)y_i \,|\, r_i y_i, r_i; \theta, \psi]\Big\} \;|\; r_i y_i, r_i; \theta^{(t)}, \psi^{(t)}\Big\}$ are the first and second terms of (3.2), respectively. The EM algorithm (Dempster et al., 1977) then proceeds as follows until the convergence criterion, ϵ, is met:

E-step: Given current estimates of the parameters, $\theta^{(t)}$ and $\psi^{(t)}$, calculate $Q\left[\theta, \psi \,|\, \theta^{(t)}, \psi^{(t)}\right]$

M-step: Maximize the Q function with respect to θ and ψ based on the expression from the E-step to obtain $\theta^{(t+1)}$ and $\psi^{(t+1)}$, and let $\theta^{(t)} = \theta^{(t+1)}$ and $\psi^{(t)} = \psi^{(t+1)}$

By Jensen's inequality, $H\left[\theta^{(t+1)}, \psi^{(t+1)} \,|\, \theta^{(t)}, \psi^{(t)}\right] \leq H\left[\theta^{(t)}, \psi^{(t)} \,|\, \theta^{(t)}, \psi^{(t)}\right]$ (Dempster et al., 1977). Due to the M-step, $Q\left[\theta^{(t+1)}, \psi^{(t+1)} \,|\, \theta^{(t)}, \psi^{(t)}\right] \geq Q\left[\theta^{(t)}, \psi^{(t)} \,|\, \theta^{(t)}, \psi^{(t)}\right]$, and thus the algorithm guarantees $L[\theta^{(t+1)}, \psi^{(t+1)}; y_{obs}, r] \geq L[\theta^{(t)}, \psi^{(t)}; y_{obs}, r]$ given the form of (3.2). As a result, the algorithm in general converges to a maximum of the likelihood (the MLE or potentially a local maximum if there are multiple maxima).

Often, sufficient statistics are updated in the EM algorithm rather than the actual data points. Without loss of generality, assume data follow a distribution belonging to the exponential family and that the pdf is in canonical form, so that

$$f(y_i;\theta) \propto exp[\theta S(y_i) + h(y_i) + b(\theta)],$$

and similar for $f(r_i \,|\, y_i;\psi)$. Then,

$$Q = \sum_{i=1}^{n} E\Big\{ln\Big\{exp[\theta S_1(y_i) + h(y_i) + b(\theta)]\Big\} + ln\Big\{exp[\psi S_2(y_i, r_i) + g(y_i, r_i) + c(\psi)]\Big\} \;|\; r_i y_i, r_i;\, \theta^{(t)}, \psi^{(t)}\Big\}$$

$$= \sum_{i=1}^{n} E\Big[\theta S_1(y_i) + h(y_i) + b(\theta) + \psi S_2(y_i, r_i) + g(y_i, r_i) + c(\psi) \mid r_i y_i, r_i;\, \theta^{(t)}, \psi^{(t)}\Big].$$

Note in the above expression of the Q function, θ and ψ depend on the data only through the sufficient statistics S_1 and S_2, so that these are the terms updated during the EM algorithm:

$$\textit{E-step}\text{: } \mathbf{S}^{(t+1)} = E\left[\mathbf{S} \mid r_i y_i, r_i; \theta^{(t)}, \psi^{(t)}\right]$$
$$\textit{M-step}\text{: } \left\{\theta^{(t+1)}, \psi^{(t+1)}\right\} = \underset{\theta,\psi}{\operatorname{argmax}}\, Q\left[\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right] = u\Big[\mathbf{S}^{(t+1)}\Big]$$

where $u(\cdot)$ is the function of the sufficient statistics that provides the MLE given the distribution the data are assumed to follow.

3.1.2 The general EM algorithm when the missing data mechanism is ignorable

If data are 1) MCAR or 2) MAR and θ and ψ are distinct ($\mathbf{\Omega}_{\theta,\psi} = \Omega_\theta \times \Omega_\psi$), the missing data mechanism may be ignored since

$$L_{full}(\theta, \psi; y_{obs}, r) = L_{ign}(\theta; y_{obs}) f(r \mid y_{obs}; \psi),$$

where $L_{ign}(\theta; y_{obs}) \propto f(y_{obs}; \theta)$ is referred to as the *ignorable likelihood* (Little and Rubin, 2002). Since the missingness does not depend on what was not observed, $f(r \mid y_{obs}; \psi)$ is known and thus only $L(\theta; y_{obs})$ needs to be considered.

Analogous to the derivations in Section 3.1.1,

$$\ell(\theta; y_{obs}) \propto \sum_{i=1}^{n} \Big\{ ln\Big[f(y_i; \theta)\Big] - ln\Big\{ f[(1 - r_i) y_i \mid r_i y_i, r_i; \theta, \psi]\Big\}\Big\},$$

and the resulting Q function is

$$\sum_{i=1}^{n} E\Big\{ln\Big[f(y_i;\theta)\Big] \mid r_i y_i;\theta^{(t)}\Big\} \equiv \sum_{i=1}^{n} \int \ell(\theta;y_i) f\left[(1-r_i)y_i \,|\, r_i y_i;\theta^{(t)}\right] dy_i.$$

Here, the same iterative algorithm as in Section 3.1.1 is used, except only θ is updated.

3.1.3 The general EM algorithm when the missing data mechanism is known

In the next section, the modified EM algorithm is introduced in the regression setting. Since the modified algorithm is related to the general EM algorithm when the missing data mechanism, ψ_0, is known, a regression model will also be used for illustration in this section.

Let X be a fully-observed predictor for all $i = 1, 2, \ldots, n$ subjects and Y an outcome subject to missingness, observed only for the first $i = 1, \ldots, c$ subjects. Let R_i indicate missing data, with $r_i = 1$ if y_i is observed, 0 o.w., and define $P(R_i = 1 \,|\, y_i, x_i) = w(y_i, x_i; \psi)$. Again assume without loss of generality the complete data follow a distribution that is a member of the exponential family and that the pdf is in canonical form.

Based on Section 3.1.1, the log-likelihood function for the regression model is expressed as

$$\ell(\theta,\psi;y_{obs},r) = \sum_{i=1}^{n} \Big\{ln\Big[f(y_i \mid x_i;\theta)\Big] + ln\Big[f(r_i \mid y_i, x_i;\psi)\Big] - ln\Big\{f\Big[(1-r_i)y_i \mid r_i y_i, r_i, x_i;\theta,\psi\Big]\Big\}\Big\} \tag{3.3}$$

Now assuming ψ_0 is known, (3.3) becomes

$$\ell(\theta; y_{obs}, r, \psi_0) = \sum_{i=1}^{n} \Big\{ ln\Big[f(y_i \mid x_i; \theta) \Big] + ln\Big[f(r_i \mid y_i, x_i; \psi_0) \Big] - ln\Big\{ f\Big[(1-r_i)y_i \mid r_i y_i, r_i, x_i; \theta, \psi_0 \Big] \Big\} \Big\}$$

and the Q function,

$$\begin{aligned} Q^*\left[\theta \,|\, \theta^{(t)}\right] &= \sum_{i=1}^{n} E\Big\{ ln[f(y_i \mid x_i; \theta)] + \; ln[f(r_i \mid y_i, x_i; \psi_0)] \mid r_i y_i, r_i, x_i; \theta^{(t)}, \psi_0 \Big\} \\ &\propto \sum_{i=1}^{n} E\Big\{ ln[f(y_i \mid x_i; \theta)] \mid r_i y_i, r_i, x_i; \theta^{(t)}, \psi_0 \Big\} \end{aligned} \tag{3.4}$$

since $ln[f(r_i \,|\, y_i, x_i; \psi_0)]$ does not involve θ and is constant when ψ_0 is known. Dividing (3.4) up into completers and incompleters yields

$$\begin{aligned} Q^*\left[\theta \,|\, \theta^{(t)}\right] &\propto \sum_{i=1}^{c} ln[f(y_i \mid x_i; \theta)] + \sum_{i=c+1}^{n} E\Big\{ ln[f(y_i \mid x_i; \theta)] \mid r_i = 0, x_i; \theta^{(t)}, \psi_0 \Big\} \\ &\propto \sum_{i=1}^{c} ln[f(y_i \mid x_i; \theta)] + \sum_{i=c+1}^{n} \Big\{ \theta E\Big[S(y_i, x_i) \mid r_i = 0, x_i; \theta^{(t)}, \psi_0 \Big] + b(\theta) \Big\}. \end{aligned}$$

Therefore, to update the Q function, $E\Big[S(y_i, x_i) \mid r_i = 0, x_i; \theta^{(t)}, \psi_0 \Big]$ must be updated. In practice, this is accomplished by calculating

$$E\left[S(y_i, x_i) \mid r_i = 0, x_i; \theta^{(t)}, \psi_0 \right] = \frac{\displaystyle\int S(y_i, x_i) f\left[y_i \,|\, x_i; \theta^{(t)} \right] [1 - w(y_i, x_i; \psi_0)] dy_i}{\displaystyle\int f\left[y_i \,|\, x_i; \theta^{(t)} \right] [1 - w(y_i, x_i; \psi_0)] dy_i}.$$

3.2 METHODS

3.2.1 A modified EM algorithm: Maximum likelihood estimation without modeling or assuming the value of the missing data mechanism

To introduce the concept of the modified algorithm, consider again the regression setting, with the theory applied to contingency tables in Section 3.2.3. The modified EM algorithm begins by assuming ψ_0 is known, then shows that under certain assumptions the value is not relevant to the estimation of θ.

As shown in Section 3.1.3, $E\left[S(y_i, x_i) \mid r_i = 0, x_i; \theta^{(t)}, \psi_0\right]$ needs to be maximized. Note

$$\begin{aligned} E\left[S(y_i, x_i) \mid x_i; \theta^{(t)}, \psi_0\right] = \; & E\left[S(y_i, x_i) \mid r_i = 1, x_i; \theta^{(t)}, \psi_0\right] P\left[R_i = 1 \mid x_i; \theta^{(t)}, \psi_0\right] + \\ & E\left[S(y_i, x_i) \mid r_i = 0, x_i; \theta^{(t)}, \psi_0\right] P\left[R_i = 0 \mid x_i; \theta^{(t)}, \psi_0\right] \end{aligned}$$

$$\begin{aligned} & \Rightarrow E\left[S(y_i, x_i) \mid r_i = 0, x_i; \theta^{(t)}, \psi_0\right] \\ & = \frac{E\left[S(y_i, x_i) \mid x_i; \theta^{(t)}, \psi_0\right] - E\left[S(y_i, x_i) \mid r_i = 1, x_i; \theta^{(t)}, \psi_0\right] P\left[R_i = 1 \mid x_i; \theta^{(t)}, \psi_0\right]}{1 - P\left[R_i = 1 \mid x_i; \theta^{(t)}, \psi_0\right]}. \end{aligned}$$

All terms on the RHS other than $E\left[S(y_i, x_i) \mid x_i; \theta^{(t)}, \psi_0\right]$ may be estimated empirically (since they are based on the fully-observed data), so only this term needs to be updated in the M-step. Note this expression depends only on X and the current estimate of θ, and subsequently the value of ψ_0 is irrelevant. Specifically,

$$\begin{aligned} & \hat{E}\left[S(y_i, x_i) \mid r_i = 0, x_i; \theta^{(t)}, \psi_0\right] \\ & \quad = \frac{E\left[S(y_i, x_i) \mid x_i; \theta^{(t)}\right] - \hat{E}\left[S(y_i, x_i) \mid r_i = 1, x_i\right] \hat{P}(R_i = 1 \mid x_i)}{1 - \hat{P}(R_i = 1 \mid x_i)}, \end{aligned} \tag{3.5}$$

where the estimate of θ in the initial E-step is assumed consistent.

This requirement of consistency is due to the replacement of some terms by their empirical estimates in the modified EM. Note these estimates will be consistent, as they are based on the fully-observed data. As a result, a sufficient condition for this replacement to be valid is that the current estimate of θ is consistent. This is achieved by beginning the modified EM algorithm with a consistent initial estimate of θ.

Obtaining these estimates may be considered more reasonable than assuming the model structure for θ and ψ, or the value of ψ_0, as these can never be truly known. Given a consistent initial estimate, the use of the EM algorithm based on the form of (3.5) will then result in a consistent estimate of θ that is presumably more efficient than the initial estimate.

3.2.2 The general EM algorithm in the contingency table setting

Consider a data set with discrete variables $W = (X, Y, Z)$, indexed by i, j, k; $i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K$, where, for simplicity, (X, Y) is fully-observed and Z is subject to missingness. Let $R_a, a = 1, 2, ..., n$, be the missing data indicator such that $r_a = 1$ if z_a is observed, 0 otherwise, and $v(w_a; \psi) = P(R_a = 1 \,|\, w_a)$. The observed data are $\{c_{ijk}, m_{ij\cdot}\}$, where the c_{ijk} represent the fully-classified table and the $m_{ij\cdot}$ the partially-classified when values of Z are missing. Let the first $c = \sum_{ijk} c_{ijk}$ observations be the number of complete cases and the next $m = \sum_{ij} m_{ij\cdot}$ be the number of incomplete cases so that $n = c + m$. Of interest is estimating $\boldsymbol{\pi} = \{\pi_{ijk} = P(X = i, Y = j, Z = k)\}$.

In the discrete data (multinomial) case, the sufficient statistics if all data are

fully-observed are

$$\mathbf{S} = \Bigg\{ n_{ijk} = c_{ijk} + m_{ijk} = \sum_{a=1}^{n=c+m} I\{x_a = i, y_a = j, z_a = k\} \,|\, i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K \Bigg\}.$$

Then, when data are missing, let $\boldsymbol{\pi}^{(t)}$ be the current estimate of $\boldsymbol{\pi}$. As derived previously, and with the current notation,

$$Q = \sum_{a=1}^{n} E\Big\{ ln\Big[f(w_a; \boldsymbol{\pi}) \Big] + ln\Big[f(r_a \,|\, w_a; \psi) \Big] \mid x_a, y_a, r_a z_a, r_a, \boldsymbol{\pi}^{(t)}, \psi^{(t)} \Big\}.$$

Let S_1 be the sufficient statistic associated with $f(w_a; \boldsymbol{\pi})$ and S_2 with $f(r_a \,|\, w_a; \psi)$, so that

$$\begin{aligned} Q &= \sum_{a=1}^{n} E\Big\{ ln\Big\{ exp\Big[\boldsymbol{\pi} S_1(w_a) + h(w_a) + b(\boldsymbol{\pi}) \Big] \Big\} + \\ &\qquad ln\Big\{ exp\Big[\psi S_2(w_a, r_a) + g(w_a, r_a) + c(\psi) \Big] \Big\} \mid x_a, y_a, r_a z_a, r_a, \boldsymbol{\pi}^{(t)}, \psi^{(t)} \Big\} \\ &= \sum_{a=1}^{n} E\Big[\boldsymbol{\pi} S_1(w_a) + h(w_a) + b(\boldsymbol{\pi}) + \psi S_2(w_a, r_a) + g(w_a, r_a) + \\ &\qquad c(\psi) \mid x_a, y_a, r_a z_a, r_a, \boldsymbol{\pi}^{(t)}, \psi^{(t)} \Big] \end{aligned}$$

Then,

$$\begin{aligned} S_{1,ijk}^{(t+1)} &= \sum_{a=1}^{c} I\{x_a = i, y_a = j, z_a = k\} + \\ &\qquad \sum_{a=c+1}^{n} I\{x_a = i, y_a = j\} E\left[I\{z_a = k\} \,|\, x_a = i, y_a = j, r_a = 0; \boldsymbol{\pi}^{(t)} \right] \\ &= c_{ijk} + \sum_{a=c+1}^{n} I\{x_a = i, y_a = j\} E\left[I\{z_a = k\} \,|\, x_a = i, y_a = j, r_a = 0; \boldsymbol{\pi}^{(t)} \right] \end{aligned}$$

and

$$S_{2,ijk}^{(t+1)} = \sum_{a=1}^{c} S(x_a = i, y_a = j, z_a = k, r_a) + \sum_{a=c+1}^{n} I\{x_a = i, y_a = j\} E\left[S(z_a = k, r_a) \,|\, x_a = i, y_a = j, r_a = 0; \boldsymbol{\pi}^{(t)}, \psi^{(t)}\right].$$

S_1 and S_2 are iteratively updated in the EM algorithm until convergence.

3.2.3 Implementation of the modified EM algorithm in the contingency table setting

In order to utilize the modified EM algorithm, assume there exists a representative external data set with $W = (X, Y, Z)$ observed for all of its n^E subjects, from which at least consistent initial estimates of the parameters, and possibly the entirety of the data, may be obtained. To differentiate the external data from that subject to missingness, define $n^M = c + m$ and $n = n^E + n^M$ from this point forward. For the following derivations, assume only the initial estimates are available from the external data. As in Section 3.1.3 and assuming ψ_0 is known, the log-likelihood is

$$\ell(\boldsymbol{\pi}; w_{obs}, r, \psi_0) = \sum_{a=1}^{n^M} \{ln[f(w_a; \boldsymbol{\pi})] + ln\left[f(r_a \mid w_a; \psi_0)\right] - \ln\left\{f\left[(1 - r_a)z_a \mid x_a, y_a, r_a z_a, r_a; \boldsymbol{\pi}, \psi_0\right]\right\}\},$$

and the Q function is

$$Q^*\left[\boldsymbol{\pi} \,|\, \boldsymbol{\pi}^{(t)}\right] \propto \sum_{a=1}^{c} ln\left[f(w_a; \boldsymbol{\pi})\right] + \sum_{a=c+1}^{n^M} \left\{\boldsymbol{\pi} E\left[S(w_a) \mid x_a, y_a, r_a = 0; \boldsymbol{\pi}^{(t)}, \psi_0\right] + b(\boldsymbol{\pi})\right\}. \tag{3.6}$$

As in Section 3.2.1, in order to update Q, update

$$\frac{E\left[S(w_a) \mid x_a, y_a; \boldsymbol{\pi}^{(t)}\right] - \hat{E}\left[S(w_a) \mid x_a, y_a, r_a = 1\right] \hat{P}(R_a = 1 \mid x_a, y_a)}{1 - \hat{P}(R_a = 1 \mid x_a, y_a)}, \tag{3.7}$$

which requires updating $E\left[S(w_a) \mid x_a, y_a, \boldsymbol{\pi}^{(t)}\right]$, as all other values are estimated from the observed data. Specifically, given the form of the sufficient statistics in Section 3.2.2,

$$\begin{aligned}
Q^*\left[\boldsymbol{\pi} \mid \boldsymbol{\pi}^{(t)}\right] &= S_{ijk}^{(t+1)} \\
&= \sum_{a=1}^{c} I\{x_a = i, y_a = j, z_a = k\} + \sum_{a=c+1}^{n^M} I\{x_a = i, y_a = j\} \times \\
&\quad E\left[I\{z_a = k\} \mid x_a = i, y_a = j, c_{ijk}, m_{ij\cdot}, r_a = 0; \boldsymbol{\pi}^{(t)}, \psi_0\right] \\
&= c_{ijk} + \sum_{a=c+1}^{n^M} I\{x_a = i, y_a = j\} \times \\
&\quad E\left[I\{z_a = k\} \mid x_a = i, y_a = j, c_{ijk}, m_{ij\cdot}, r_a = 0; \boldsymbol{\pi}^{(t)}, \psi_0\right],
\end{aligned}$$

where

$$\begin{aligned}
&E\left[I\{z_a = k\} \mid x_a = i, y_a = j, c_{ijk}, m_{ij\cdot}, r_a = 0; \boldsymbol{\pi}^{(t)}, \psi_0\right] \\
&\quad = \Big\{ E\left[I\{z_a = k\} \mid x_a = i, y_a = j; \boldsymbol{\pi}^{(t)}\right] \; - \\
&\qquad \hat{E}\left(I\{z_a = k\} \mid x_a = i, y_a = j, c_{ijk}, r_a = 1\right) \times \\
&\qquad \hat{P}(R_a = 1 \mid x_a = i, y_a = j) \Big\} \Big/ \\
&\qquad \Big[1 - \hat{P}(R_a = 1 \mid x_a = i, y_a = j)\Big] \text{ based on the form of (3.7)} \\
&\quad = \frac{\left[\frac{\pi_{ijk}^{(t)}}{\pi_{ij\cdot}^{(t)}} - \frac{c_{ijk}}{c_{ij\cdot}} \frac{c_{ij\cdot}}{c_{ij\cdot} + m_{ij\cdot}}\right]}{\left(1 - \frac{c_{ij\cdot}}{c_{ij\cdot} + m_{ij\cdot}}\right)}
\end{aligned}$$

$$= \frac{c_{ij\cdot} + m_{ij\cdot}}{m_{ij\cdot}} \frac{\pi^{(t)}_{ijk}}{\pi^{(t)}_{ij\cdot}} - \frac{c_{ijk}}{m_{ij\cdot}} \tag{3.8}$$

$$\begin{aligned}
\Rightarrow S^{(t+1)}_{ijk} &= c_{ijk} + m_{ij\cdot} \left[\frac{c_{ij\cdot} + m_{ij\cdot}}{m_{ij\cdot}} \frac{\pi^{(t)}_{ijk}}{\pi^{(t)}_{ij\cdot}} - \frac{c_{ijk}}{m_{ij\cdot}} \right] \\
&= c_{ijk} + (c_{ij\cdot} + m_{ij\cdot}) \frac{\pi^{(t)}_{ijk}}{\pi^{(t)}_{ij\cdot}} - c_{ijk} \\
&= (c_{ij\cdot} + m_{ij\cdot}) \frac{\pi^{(t)}_{ijk}}{\pi^{(t)}_{ij\cdot}} \\
&= n^{M}_{ij\cdot} \frac{\pi^{(t)}_{ijk}}{\pi^{(t)}_{ij\cdot}}.
\end{aligned} \tag{3.9}$$

If one were able to obtain the actual data and not just initial estimates, n^{E}_{ijk} would be added to the RHS of (3.9).

Based on the form of (3.9), one will note the only addition of information from the data set subject to missingness is through the term $n^{M}_{ij\cdot}$. In other words, only the (X, Y) margin is updated, while any values of Z, even if they had been observed, are ignored. As a result, the modified EM in the context of discrete data simplifies to using the general EM algorithm as follows:

1. Calculate consistent initial estimates from the external data.
2. Delete all values of Z (the variable subject to missingness) amongst completers in the data set subject to missingness.
3. Conduct the standard EM algorithm using the data from (2), starting with the consistent initial estimates from (1).

This outline is assuming the external data itself is not available for use, but rather just its estimates. If instead it were, after step (2), one would concatenate the external data and that from (2), such that in the resulting data set, X and Y are

fully-observed, but Z is only observed in the portion of the data associated with the external data.

Given this simplification in the context of discrete data, the modified EM algorithm is not iterative under every model structure, but instead may converge in two steps, which is illustrated in Section 3.2.4.2. This is due to the fact that after the (X, Y) margin, $n_{ij\cdot}^{M}$, is incorporated as in (3.9), no more information can be gained by further iterations and thus the estimates stabilize. As such, the method *does* increase efficiency by incorporating information on X and Y, but in some instances, this is a closed-form, not iterative, process.

Although this approach is naïve in that it deletes information (values of Z in the data subject to missingness), it is this deletion that allows the modified EM algorithm to produce unbiased estimates while increasing efficiency. The only way information on Z is included in the final estimates is through the external data, which is assumed to provide consistent estimation. As such, if bias is of particular concern in a given analysis, this approach may be favorable when contrasted with making assumptions about the missing data mechanism, as is required in the general EM.

In the following section, the modified EM algorithm is considered under three different discrete model structures. Based on whether or not the $n_{ij\cdot}$ are in the set of sufficient statistics, a general conclusion is drawn regarding the convergence rate of the modified EM. Additionally, this criterion determines when consistent initial estimates (in conjunction with a data set subject to missingness) are sufficient for an increase in efficiency, compared to model structures that additionally require the use of the external data itself in order to provide such an increase.

3.2.4 Applications in contingency table analyses of missing data

In the following sections, three types of discrete data model structures are considered: a conditional independence model, a saturated model and a three-way table without a three-way interaction (i.e., a two-way interaction model). These structures are divided into those that do and do not include the $n_{ij\cdot}$ in their set of sufficient statistics.

3.2.4.1 Models without $n_{ij\cdot}$ in the set of sufficient statistics In the model that represents conditional independence of X and Y given Z, $(X \perp Y) \,|\, Z \Rightarrow (X \,|\, Z) \perp (Y \,|\, Z)$, so that $\pi_{ijk} = \pi_{i\,|\,k}\pi_{j\,|\,k}\pi_{\cdot\cdot k}$. In this case, the sufficient statistics are $n_{i\,|\,k}$, $n_{j\,|\,k}$ and $n_{\cdot\cdot k}$, and thus

$$\pi_{i\,|\,k}^{(t+1)} = \frac{S_{i\cdot k}^{(t+1)}}{S_{\cdot\cdot k}^{(t+1)}},\ \pi_{j\,|\,k}^{(t+1)} = \frac{S_{\cdot jk}^{(t+1)}}{S_{\cdot\cdot k}^{(t+1)}} \text{ and } \pi_{\cdot\cdot k}^{(t+1)} = \frac{S_{\cdot\cdot k}^{(t+1)}}{n^M} \tag{3.10}$$

$$\Rightarrow \pi_{ijk}^{(t+1)} = \pi_{i\,|\,k}^{(t+1)}\pi_{j\,|\,k}^{(t+1)}\pi_{\cdot\cdot k}^{(t+1)} \text{ and } \pi_{ij\cdot}^{(t+1)} = \sum_k \pi_{ijk}^{(t+1)}, \tag{3.11}$$

If n_{ijk}^E were available and used in (3.9), the denominator of $\pi_{\cdot\cdot k}^{(t+1)}$ would be n instead of n^M.

The use of the modified EM algorithm under this model structure is iterative (i.e., it does not converge in two iterations). Additionally, this structure does *not* benefit from consistent initial estimates alone, but rather requires the external data as well, as is shown via simulation in Section 3.3.2.1.

Note here that all sufficient statistics involve Z, either conditionally or marginally. Thus, the (X, Y) marginal information from the data set subject to missingness cannot improve the initial estimates alone. In order to increase efficiency here, there must be an increase in sample size − namely, the external data and that subject to missingness must be combined. The modified EM algorithm allows for this, without

inducing bias into the final estimates, despite the fact that estimates based on the data set subject to missingness alone could have been biased.

3.2.4.2 Models with $n_{ij\cdot}$ in the set of sufficient statistics Consider first the saturated model, where no form of independence or interaction amongst variables is assumed. Thus, $\hat{\pi}_{ijk} = \frac{n_{ijk}}{n}$ when data are fully-observed, with sufficient statistics n_{ijk}. Then, since $\pi_{ijk} = \pi_{ij\cdot}\pi_{k\,|\,ij}$ in general, updating π_{ijk} is equivalent to updating $\pi_{ij\cdot}\pi_{k\,|\,ij}$. Given the sufficient statistics here are $n_{ij\cdot}$ and $n_{k\,|\,ij}$, the information from the $n^M_{ij\cdot}$ is incorporated directly. Because of this, the proposed EM algorithm will stabilize within two iterations:

$$
\begin{aligned}
\pi^{(0)}_{ijk} &= \frac{n^E_{ijk}}{n^E} \text{ and } \pi^{(0)}_{ij\cdot} = \frac{n^E_{ij\cdot}}{n^E} \\
\Rightarrow S^{(1)}_{ijk} &= n^M_{ij\cdot}\frac{\pi^{(0)}_{ijk}}{\pi^{(0)}_{ij\cdot}} \\
&= (n_{ij\cdot} - n^E_{ij\cdot})\frac{n^E_{ijk}}{n^E_{ij\cdot}} \\
&= \frac{n_{ij\cdot}n^E_{ijk}}{n^E_{ij\cdot}} - \frac{n^E_{ij\cdot}n^E_{ijk}}{n^E_{ij\cdot}} \\
&= \frac{n_{ij\cdot}n^E_{ijk}}{n^E_{ij\cdot}} - n^E_{ijk} \\
&= n_{ij\cdot}\frac{\pi^{(0)}_{ijk}}{\pi^{(0)}_{ij\cdot}} - n^E_{ijk} \\
\Rightarrow \pi^{(1)}_{ijk} &= \frac{S^{(1)}_{ijk}}{n^M} = \frac{n_{ij\cdot}}{n^M}\frac{\pi^{(0)}_{ijk}}{\pi^{(0)}_{ij\cdot}} - \frac{n^E_{ijk}}{n^M} \\
\Rightarrow \pi^{(1)}_{ij\cdot} &= \sum_k \pi^{(1)}_{ijk} = \frac{n_{ij\cdot} - n^E_{ij\cdot}}{n^M} = \frac{n^M_{ij\cdot}}{n^M}
\end{aligned}
$$

$$\Rightarrow S_{ijk}^{(2)} = n_{ij.}^{M} \frac{\left[\frac{n_{ij.}}{n^M}\frac{\pi_{ijk}^{(0)}}{\pi_{ij.}^{(0)}} - \frac{n_{ijk}^{E}}{n^M}\right]}{\left(\frac{n_{ij.}^{M}}{n^M}\right)}$$

$$= S_{ijk}^{(1)},$$

which indicates $\pi_{ijk}^{(t+1)} = \pi_{ijk}^{(t)} \; \forall \; t > 0$.

The following expression shows how the data set subject to missingness is incorporated into the initial estimate of $\boldsymbol{\pi}$:

$$\begin{aligned}
\pi_{ijk}^{(1)} &= \frac{n_{ij.}}{n^M}\frac{\pi_{ijk}^{(0)}}{\pi_{ij.}^{(0)}} - \frac{n_{ijk}^{E}}{n^M} \\
&= \frac{n_{ij.}}{n^M}\frac{\pi_{ijk}^{(0)}}{\left(\frac{n_{ij.}^{E}}{n^E}\right)} - \frac{n_{ijk}^{E}\left(\frac{n_{ij.}^{E}}{n^E}\right)}{n^M\left(\frac{n_{ij.}^{E}}{n^E}\right)} \\
&= \frac{\pi_{ijk}^{(0)} n^E (n_{ij.} - n_{ij.}^{E})}{n^M n_{ij.}^{E}} \\
&= \frac{\pi_{ijk}^{(0)} n^E n_{ij.}^{M}}{n^M n_{ij.}^{E}}
\end{aligned}$$

Thus, the original estimate based on the external data, $\pi_{ijk}^{(0)}$, is augmented by the factor $\frac{n^E n_{ij.}^{M}}{n^M n_{ij.}^{E}}$, giving a final estimate in closed-form that includes information from the data subject to missingness, but only through the (X, Y) margin.

Due to the fact that information from the $n_{ij.}^{M}$ is incorporated directly into the sufficient statistics, efficiency is increased given the external estimates alone for a sample size only slightly larger than that of the external data (e.g., $n^M = 320$ vs. $n^E = 300$). This fact is not shown via simulation for this model structure, but results are analogous to those under the two-way interaction model, found in Section 3.3.2.2.

Now consider the two-way interaction model (three-way contingency table with no three-way interaction), where the sufficient statistics are $n_{ij.}$, $n_{i.k}$, $n_{.jk}$. In order

to correctly model the interactions between X, Y and Z, a log-linear model must be used, where the cell counts of the contingency table are Poisson-distributed with expected frequencies $\mu_{ijk} = n\pi_{ijk}$ (Agresti, 2002). The model of interest is parameterized as

$$ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{3.12}$$

As there is no closed-form for the π_{ijk} associated with this model, *iterative proportional fitting* (IPF) is utilized to obtain estimates of $\boldsymbol{\pi}$ (Agresti, 2002). For this model structure, IPF has three steps (illustrated for obtaining estimates from the external data):

$$\mu_{ijk}^{(1)} = \mu_{ijk}^{(0)} \frac{n_{ij\cdot}^E}{\mu_{ij\cdot}^{(0)}}, \quad \mu_{ijk}^{(2)} = \mu_{ijk}^{(1)} \frac{n_{i\cdot k}^E}{\mu_{i\cdot k}^{(1)}}, \quad \mu_{ijk}^{(3)} = \mu_{ijk}^{(2)} \frac{n_{\cdot jk}^E}{\mu_{\cdot jk}^{(2)}}. \tag{3.13}$$

The initial estimates, $\mu_{ijk}^{(0)}$ and $\mu_{ij\cdot}^{(0)}$, may be set trivially to 1. After one cycle (i.e., all three steps) is complete, $\mu_{ij\cdot}^{(3)}$ is compared with $n_{ij\cdot}^E$, $\mu_{i\cdot k}^{(3)}$ with $n_{i\cdot k}^E$ and $\mu_{\cdot jk}^{(3)}$ with $n_{\cdot jk}^E$. If convergence is not met, $\mu_{ijk}^{(0)} \overset{set}{=} \mu_{ijk}^{(3)}$ and cycles are continued until convergence. Denoting the estimates of the μ_{ijk} resulting from IPF as $\hat{\mu}_{ijk}$, estimates of the elements of $\boldsymbol{\pi}$ are given by $\hat{\pi}_{ijk} = \dfrac{\hat{\mu}_{ijk}}{\sum_{i,j,k} \hat{\mu}_{ijk}}$.

During the modified EM algorithm, one would use $S_{ij\cdot}^{(t+1)}$, $S_{i\cdot k}^{(t+1)}$ and $S_{\cdot jk}^{(t+1)}$ in place of $n_{ij\cdot}^E$, $n_{i\cdot k}^E$ and $n_{\cdot jk}^E$ in (3.13). Therefore, estimation of the π_{ijk} depends not on $S_{ijk}^{(t+1)}$, but rather on these marginal statistics. As a result, if these statistics stabilize, the estimates of the μ_{ijk} and thus π_{ijk} no longer update. The following derivations show how the modified EM converges in two iterations under this model structure: Note in general that $S_{ijk}^{(t+1)} = n_{ij\cdot}^M \frac{\pi_{ijk}^{(t)}}{\pi_{ij\cdot}^{(t)}}$ from (3.9), which implies $S_{ijk}^{(1)} = n_{ij\cdot}^M \frac{\pi_{ijk}^{(0)}}{\pi_{ij\cdot}^{(0)}}$, where

$\pi^{(0)}_{ijk}$ and $\pi^{(0)}_{ij\cdot}$ are obtained from the external data via IPF. Then,

$$S^{(1)}_{ij\cdot} = \sum_k S^{(1)}_{ijk} = \sum_k n^M_{ij\cdot} \frac{\pi^{(0)}_{ijk}}{\pi^{(0)}_{ij\cdot}} = n^M_{ij\cdot} \ \forall \ t.$$

Subsequently, $\pi^{(1)}_{ij\cdot} = \frac{S^{(1)}_{ij\cdot}}{n^M} = \frac{n^M_{ij\cdot}}{n^M}$, which holds true $\forall\ t > 0$. Then,

$$S^{(2)}_{i\cdot k} = \sum_j S^{(2)}_{ijk} = \sum_j n^M_{ij\cdot} \frac{\pi^{(1)}_{ijk}}{\pi^{(1)}_{ij\cdot}} = \sum_j n^M_{ij\cdot} \frac{\pi^{(1)}_{ijk}}{\left(\frac{n^M_{ij\cdot}}{n^M}\right)} = n^M \sum_j \pi^{(1)}_{ijk} = S^{(1)}_{i\cdot k},$$

and similar for $S^{(2)}_{\cdot jk}$. Thus, by the second iteration, the marginal sufficient statistics stabilize and the estimates of $\boldsymbol{\pi}$ will not benefit from the iterative method. This is another instance where the (X, Y) margin from the data set subject to missingness is incorporated during the first iteration, providing some improvement in efficiency.

As with the saturated model, since $n_{ij\cdot}$ are sufficient statistics here, efficiency is increased when only the initial estimates from the external data are available. These results are reflected in a simulation study in Section 3.3.2.2.

3.2.5 The role of sufficient statistics in the modified EM algorithm

As discussed above, whether or not $n_{ij\cdot}$ are in the set of sufficient statistics given the assumed model structure is the determining factor in whether or not the external data itself is required in order increase efficiency through the modified EM algorithm. If $n_{ij\cdot}$ are sufficient statistics, only the initial estimates are required; if not, the external data set itself is additionally needed. In the latter case, the increase in efficiency is through a sheer increase in sample size. However, the valuable aspect of the modified EM algorithm is that data that would otherwise provide biased estimates may be utilized without actually inducing bias.

3.3 SIMULATION STUDIES

3.3.1 Defining the data structure

Below are the procedures used to simulate data according to two model structures: conditional independence and two-way interaction (three-way contingency table with no three-way interaction).

3.3.1.1 Three-way contingency table with conditional independence The following steps are used to derive the joint distribution of three binary variables, X, Y and Z, under the assumption of $(X \perp Y) \,|\, Z$. From this, a fully-observed data set is created, as well as that subject to missingness under either MCAR, MAR or MNAR:

Step 1: Derive the joint distribution of X, Y and Z:

(a) For binary $X, Y, Z \in \{1, 2\}$, fix the values of $P(Z = 2), P(X = 2 \,|\, Z = 1), P(X = 2 \,|\, Z = 2), P(Y = 2 \,|\, Z = 1)$ and $P(Y = 2 \,|\, Z = 2)$.

(b) Find the joint distribution of X, Y and Z, $\boldsymbol{\pi} = (\pi_{111}, \pi_{121}, \pi_{211}, \pi_{221}, \pi_{112}, \pi_{122}, \pi_{212}, \pi_{222})$, using the fact that $(X \perp Y) \,|\, Z \equiv (X \,|\, Z) \perp (Y \,|\, Z) \Rightarrow P(X \,|\, Z)P(Y \,|\, Z) = P[(X, Y) \,|\, Z] \Rightarrow P(X \,|\, Z)P(Y \,|\, Z)P(Z) = P(X, Y, Z)$.

Step 2: Create a fully-observed data set by sampling n^E observations from the multinomial distribution from *Step 1* and let $n^E_{ijk} = \sum_{a \in \mathcal{D}_E} I\{X_a = i, Y_a = j, Z_a = k\}$, where $\mathcal{D}_E$ represents the set of indices for subjects in the external data set.

Step 3: Create a data set subject to missingness:

(a) Sample n^M observations from the distribution determined in *Step 1* and let $n^M_{ijk} = \sum_{a \in \mathcal{D}_M} I\{X_a = i, Y_a = j, Z_a = k\}$, where $\mathcal{D}_M$ represents the set

of indices for subjects in the data set subject to missingness.

(b) For data MCAR, let $m_{ijk} \sim BIN(n^M_{ijk}, \psi)\ \forall\ i, j, k$; else go to *Step 3*$(c)$.

(c) For data MAR, let $m_{ijk} \sim BIN(n^M_{ijk}, \psi_{ij\cdot})\ \forall\ i, j, k$; else go to *Step 3*$(d)$.

(d) For data MNAR, let $m_{ijk} \sim BIN(n^M_{ijk}, \psi_{\cdot\cdot k})\ \forall\ i, j, k$.

(e) The observed data is then $c_{ijk} = n^M_{ijk} - m_{ijk}$ and $m_{ij\cdot} = m_{ij1} + m_{ij2}\ \forall\ i, j, k$.

3.3.1.2 Three-way contingency table with no three-way interaction As in Section 3.3.1.1, the following steps are used to simulate a fully-observed data set, as well as that subject to missingness for the model that includes all two-way interactions between X, Y and Z, but not the three-way interaction.

Step 1: Using the form of (3.12), choose λ^X_i, λ^Y_j, λ^Z_k, λ^{XY}_{ij}, λ^{XZ}_{ik} and λ^{YZ}_{jk} such that all $\pi_{ijk} = \dfrac{\mu_{ijk}}{\sum\limits_{i,j,k} \mu_{ijk}} > 0.05$ so as to avoid sparse cells.

Step 2: Create a fully-observed data set by sampling n^E observations from the multinomial distribution from *Step 1* and let $n^E_{ijk} = \sum\limits_{a \in \mathcal{D}_E} I\{X_a = i, Y_a = j, Z_a = k\}$, where $\mathcal{D}_E$ represents the set of indices for subjects in the external data set.

Step 3: Create a data set subject to missingness:

(a) Sample n^M observations from the distribution determined in *Step 1* and let $n^M_{ijk} = \sum\limits_{a \in \mathcal{D}_M} I\{X_a = i, Y_a = j, Z_a = k\}$, where $\mathcal{D}_M$ represents the set of indices for subjects in the data set subject to missingness.

(b) For data MCAR, let $m_{ijk} \sim BIN(n^M_{ijk}, \psi)\ \forall\ i, j, k$; else go to *Step 3*$(c)$.

(c) For data MAR, let $m_{ijk} \sim BIN(n^M_{ijk}, \psi_{ij\cdot})\ \forall\ i, j, k$; else go to *Step 3*$(d)$.

(d) For data MNAR, let $m_{ijk} \sim BIN(n^M_{ijk}, \psi_{..k}) \; \forall \; i, j, k.$

(e) The observed data is then $c_{ijk} = n^M_{ijk} - m_{ijk}$ and $m_{ij.} = m_{ij1} + m_{ij2} \; \forall \; i, j, k.$

3.3.2 Simulation of missing data using the modified EM algorithm

This section outlines two algorithms that assess the empirical bias and standard deviation of estimates under the assumption of data missing not at random in the following scenarios: 1) use of the external data alone, 2) the complete-case analysis based on the data set subject to missingness, 3) use of the modified EM algorithm with consistent initial estimates from the external data only and 4) use of the modified EM algorithm with both consistent initial estimates and the external data itself. *Algorithm 1* provides these results for the conditional independence model, while *Algorithm 2* does so for the case of the two-way interaction model.

3.3.2.1 Three-way contingency table with conditional independence

Algorithm 1: Based on the conditional independence model structure (Section 3.3.1.1), calculate the empirical bias and standard deviation of estimates under data assumed MNAR for the following cases: 1) use of the external data alone, 2) the complete-case analysis based on the data set subject to missingness, 3) use of the modified EM algorithm with consistent initial estimates from the external data only, and 4) use of the modified EM algorithm with both consistent initial estimates and the external data itself.

Step 1: For $d = 1, \ldots, D$:

(a) *Simulate data.* For fixed n^E, n^M and distribution of (X, Y, Z) as determined in *Step 1* of Section 3.3.1.1, conduct steps *2-3* of Section 3.3.1.1, using option (d) in *Step 3* (i.e. simulate external data and that missing not at random).

(b) *Estimates from external data only.* Calculate estimates of the joint distribution of (X, Y, Z) under the assumption $(X \perp Y) \,|\, Z$ based on the external data only: $\pi_{ijk}^{(0)(d)} = \pi_{i\,|\,k}^{(0)(d)} \, \pi_{j\,|\,k}^{(0)(d)} \pi_{\cdot\cdot k}^{(0)(d)} \; \forall \; i, j, k$, where $\pi_{i\,|\,k}^{(0)(d)} = \frac{n_{i\cdot k}^E}{n_{\cdot\cdot k}^E}$, $\pi_{j\,|\,k}^{(0)(d)} = \frac{n_{\cdot jk}^E}{n_{\cdot\cdot k}^E}$ and $\pi_{\cdot\cdot k}^{(0)(d)} = \frac{n_{\cdot\cdot k}^E}{n^E}$.

(c) *Bias in external data estimates.* For the estimates based on the external data only, store the vector of biases as $\boldsymbol{\delta}_E^{(d)} = \boldsymbol{\pi}^{(0)(d)} - \boldsymbol{\pi}$.

(d) *Estimates from complete cases.* Calculate estimates of the joint distribution of (X, Y, Z) under the assumption $(X \perp Y) \,|\, Z$ based on the complete cases only: $\pi_{ijk}^{CC(d)} = \pi_{i\,|\,k}^{CC(d)} \, \pi_{j\,|\,k}^{CC(d)} \pi_{\cdot\cdot k}^{CC(d)} \; \forall \; i, j, k$, where $\pi_{i\,|\,k}^{CC(d)} = \frac{c_{i\cdot k}}{c_{\cdot\cdot k}}$, $\pi_{j\,|\,k}^{CC(d)} = \frac{c_{\cdot jk}}{c_{\cdot\cdot k}}$ and $\pi_{\cdot\cdot k}^{CC(d)} = \frac{c_{\cdot\cdot k}}{c}$.

(e) *Bias in complete case estimates.* For the estimates based on the complete cases, store the vector of biases as $\boldsymbol{\delta}_{CC}^{(d)} = \boldsymbol{\pi}^{CC(d)} - \boldsymbol{\pi}$.

(f) *Modified EM algorithm with initial estimates only.* Use the form of (3.9) to obtain estimates from the modified EM algorithm using the consistent initial estimates from the external data, $\pi_{ijk}^{(0)(d)}$: Let $t = 0$, then:

(i) Calculate $S_{ijk}^{t+1} = n_{ij\cdot} \frac{\pi_{ijk}^{(t)}}{\pi_{ij\cdot}^{(t)}}$ (Eq. (3.9)).

(ii) Calculate the quantities $\pi_{i\,|\,k}^{(t+1)}$, $\pi_{j\,|\,k}^{(t+1)}$ and $\pi_{\cdot\cdot k}^{(t+1)}$ as in (3.10).

(iii) Calculate $\pi_{ijk}^{(t+1)}$ and $\pi_{ij\cdot}^{(t+1)}$ as in (3.11).

(iv) Let $t = t + 1$.

(v) Repeat steps *1*$(f)(i)$-(iv) until $\left|\pi_{ijk}^{(t+1)} - \pi_{ijk}^{(t)}\right| < \epsilon \ \forall \ i, j, k$, where ϵ is the required level of convergence. Denote the resulting estimates after convergence as $\pi_{ijk}^{I(d)}$, where I represents the estimates based on the initial estimates only.

(g) *Bias in estimates from modified EM algorithm using consistent initial estimates only.* For the estimates from the modified EM algorithm that only uses the initial estimates from the external data, store the vector of biases as $\boldsymbol{\delta}_I^{(d)} = \boldsymbol{\pi}^{I(d)} - \boldsymbol{\pi}$.

(h) *Modified EM algorithm with external data.* Obtain estimates from the modified EM algorithm using the consistent initial estimates from the external data, as well as the external data itself: Let $t = 0$, then conduct steps *1*$(f)(i)$-(v), with the exception that $S_{ijk}^{(t+1)} = n_{ij\cdot}^{M} \frac{\pi_{ijk}^{(t)}}{\pi_{ij\cdot}^{(t)}} + n_{ijk}^{E}$ in *Step 1*$(f)(i)$. Denote the resulting estimates after convergence as $\pi_{ijk}^{A(d)}$, where A represents the estimates based on all data.

(i) *Bias in estimates from modified EM algorithm using external data.* For the estimates from the modified EM algorithm that uses the external data, store the vector of biases as $\boldsymbol{\delta}_A^{(d)} = \boldsymbol{\pi}^{A(d)} - \boldsymbol{\pi}$.

Step 2: Calculate the average empirical bias of the estimates of the joint distribution based on the external data only as $\frac{1}{D}\sum_{d=1}^{D} \boldsymbol{\delta}_E^{(d)}$, and similar for all other types of estimates (complete cases, etc.).

Step 3: Calculate the standard deviation of the estimates of the joint distribution based on the external data only as $\sqrt{\frac{\sum_{d=1}^{D}\left[\boldsymbol{\pi}^{E(d)} - \bar{\boldsymbol{\pi}}^{E}\right]^2}{D-1}}$, where $\bar{\boldsymbol{\pi}}^{E}$ is the

vector of means of $\boldsymbol{\pi}^{E(d)}$. The standard deviation for the the other estimates (complete cases, etc.) follows similarly.

Table 3.1 shows the results of *Algorithm 1* (conditional independence model) for data MNAR with $\psi_{\cdot\cdot 1} = 0.55$ and $\psi_{\cdot\cdot 2} = 0.3$. In all cases, $n^E = 300$. Empirical bias and SD were calculated over $D = 2000$ iterations, with $\epsilon = 1.0 \times 10^{-8}$.

As discussed in Section 3.2.4.1, initial estimates alone are not sufficient to increase efficiency under this model structure. When $n^M = 320$ (i.e., slightly larger than n^E), the estimates from the modified EM algorithm given the initial estimates only (Table 3.1, column 6) were never more efficient than those from the external data alone. Even when the sample size was increased markedly to $n^M = 5000$, they were not more efficient (the one instance of greater efficiency is a remnant of the large sample size). This contrasts the last column of Table 3.1, in which estimates from the modified EM that included the external data were always more efficient than those based on the external data alone. Additionally, these estimates were unbiased (relative to the magnitude of the SD), despite the fact that the data were missing not at random. The complete-case analysis (column 5) shows the bias in the estimates due to this missing data mechanism.

Table 3.1: Performance of the modified EM algorithm for a conditional independence model with data MNAR. The average empirical bias (SD) for estimates of the distribution of (X, Y, Z) is reported under various estimation methods. SDs in bold represent estimates as or more efficient than those based on the external data alone.

n^M	$\boldsymbol{\pi}$	Population parameters	External data ($n^E = 300$)	Complete cases	Initial estimates from external data	Initial estimates and external data
320	π_{111}	0.231	-5.94e-04 (0.0220)	4.22e-02 (0.0291)	-1.11e-02 (0.0407)	-4.21e-04 (**0.0190**)
	π_{121}	0.099	-4.16e-04 (0.0141)	1.86e-02 (0.0204)	-1.35e-03 (0.0313)	-1.25e-04 (**0.0117**)
	π_{211}	0.189	3.24e-04 (0.0202)	3.50e-02 (0.0275)	8.21e-03 (0.0395)	-4.84e-06 (**0.0176**)
	π_{221}	0.081	4.72e-05 (0.0119)	1.36e-02 (0.0187)	6.92e-03 (0.0211)	2.03e-05 (**0.0104**)
	π_{112}	0.117	-2.34e-04 (0.0165)	-3.27e-02 (0.0280)	1.05e-02 (0.0291)	-6.91e-05 (**0.0155**)
	π_{122}	0.063	-3.67e-04 (0.0110)	-1.72e-02 (0.0201)	1.74e-03 (0.0264)	-2.51e-04 (**0.0102**)
	π_{212}	0.143	9.91e-04 (0.0183)	-3.91e-02 (0.0295)	-7.73e-03 (0.0419)	6.96e-04 (**0.0170**)
	π_{222}	0.077	2.48e-04 (0.0126)	-2.06e-02 (0.0223)	-7.24e-03 (0.0319)	1.55e-04 (**0.0113**)
5000	π_{111}	0.231	-7.35e-04 (0.0220)	4.18e-02 (0.0074)	-1.06e-03 (**0.0204**)	-2.99e-04 (**0.0150**)
	π_{121}	0.099	-3.66e-04 (0.0141)	1.80e-02 (0.0054)	-4.09e-04 (0.0143)	-6.32e-05 (**0.0099**)
	π_{211}	0.189	1.30e-04 (0.0206)	3.42e-02 (0.0069)	6.84e-04 (0.0207)	-3.85e-04 (**0.0159**)
	π_{221}	0.081	6.45e-05 (0.0120)	1.48e-02 (0.0048)	4.25e-04 (0.0121)	-6.68e-05 (**0.0088**)
	π_{112}	0.117	-3.19e-05 (0.0163)	-3.17e-02 (0.0067)	9.83e-04 (0.0181)	1.90e-04 (**0.0139**)
	π_{122}	0.063	-3.13e-04 (0.0110)	-1.73e-02 (0.0051)	2.51e-04 (0.0131)	-1.32e-04 (**0.0092**)
	π_{212}	0.143	1.07e-03 (0.0184)	-3.88e-02 (0.0075)	-5.16e-04 (0.0211)	6.03e-04 (**0.0159**)
	π_{222}	0.077	1.80e-04 (0.0131)	-2.10e-02 (0.0057)	-3.57e-06 (0.0144)	1.54e-04 (**0.0099**)

3.3.2.2 Three-way contingency table with no three-way interaction

Algorithm 2: Based on the two-way interaction model structure (Section 3.3.1.2), calculate the empirical bias and standard deviation of estimates under data assumed MNAR for the following cases: 1) use of the external data alone, 2) the complete-case analysis based on the data set subject to missingness, 3) use of the modified EM algorithm with consistent initial estimates from the external data only, and 4) use of the modified EM algorithm with both consistent initial estimates and the external data itself.

Step 1: For $d = 1, \ldots, D$:

(a) *Simulate data.* For fixed n^E, n^M and distribution of (X, Y, Z) as determined in *Step 1* of Section 3.3.1.2, conduct steps *2-3* of Section 3.3.1.2, using option (d) in *Step 3* (i.e. simulate external data and that missing not at random).

(b) *Estimates from external data only.* Calculate estimates of the joint distribution of (X, Y, Z) under the two-way interaction model based on the external data only, $\pi_{ijk}^{(0)(d)}$, using IPF:

(i) Calculate $n_{ij\cdot}^E$, $n_{i\cdot k}^E$ and $n_{\cdot jk}^E$.

(ii) Set $\mu_{ijk}^{(0)}$ and $\mu_{ij\cdot}^{(0)}$ trivially to 1.

(iii) Conduct the cycle of steps in (3.13) until convergence at level ϵ_1.

(iv) Store $\pi_{ijk}^{(0)(d)} = \dfrac{\hat{\mu}_{ijk}}{\sum_{i,j,k} \hat{\mu}_{ijk}}$, where $\hat{\mu}_{ijk}$ represents the estimates resulting from the IPF in *Step 1*$(b)(iii)$.

(c) *Bias in external data estimates.* For the estimates based on the external data only, store the vector of biases as $\boldsymbol{\delta}_E^{(d)} = \boldsymbol{\pi}^{(0)(d)} - \boldsymbol{\pi}$.

(d) *Estimates from complete cases.* To estimate $\boldsymbol{\pi}$ based on complete cases, conduct *Step 1*(b), using $c_{ij\cdot}$, $c_{i\cdot k}$ and $c_{\cdot jk}$ in place of $n^E_{ij\cdot}$, $n^E_{i\cdot k}$ and $n^E_{\cdot jk}$, and denote the estimates in *Step 1*$(b)(iv)$ as $\pi^{CC(d)}_{ijk}$ instead of $\pi^{(0)(d)}_{ijk}$.

(e) *Bias in complete case estimates.* For the estimates based on the complete cases, store the vector of biases as $\boldsymbol{\delta}^{(d)}_{CC} = \boldsymbol{\pi}^{CC(d)} - \boldsymbol{\pi}$.

(f) *Modified EM algorithm with initial estimates only.* Use the form of (3.9) and IPF to obtain estimates from the modified EM algorithm using the consistent initial estimates from the external data, $\pi^{(0)(d)}_{ijk}$: Let $t = 0$, then:

(i) Calculate $S^{(t+1)}_{ijk} = n^M_{ij\cdot}\frac{\pi^{(t)}_{ijk}}{\pi^{(t)}_{ij\cdot}}$.

(ii) Use IPF as in *Step 1*(b) to calculate $\pi^{(t+1)}_{ijk}$ using $S^{(t+1)}_{ij\cdot}$, $S^{(t+1)}_{i\cdot k}$ and $S^{(t+1)}_{\cdot jk}$ in place of $n^E_{ij\cdot}$, $n^E_{i\cdot k}$ and $n^E_{\cdot jk}$.

(iii) Let $t = t + 1$.

(iv) Repeat steps *1*$(f)(i)$-(iii) until $\left|\pi^{(t+1)}_{ijk} - \pi^{(t)}_{ijk}\right| < \epsilon_2 \ \forall\ i, j, k$, where $\epsilon_2 > \epsilon_1$ is the required level of convergence (the initial estimates, $\pi^{(0)(d)}_{ijk}$, may not provide enough precision for this step to converge at the same level, ϵ_1). Denote the resulting estimates after convergence as $\pi^{I(d)}_{ijk}$, where I represents the estimates based on the initial estimates only.

(g) *Bias in estimates from modified EM algorithm using consistent initial estimates only.* For the estimates from the modified EM algorithm that only uses the initial estimates from the external data, store the vector of biases as $\boldsymbol{\delta}^{(d)}_{I} = \boldsymbol{\pi}^{I(d)} - \boldsymbol{\pi}$.

(h) *Modified EM algorithm with external data.* Obtain estimates from the modified EM algorithm using the consistent initial estimates from the external data, as well as the external data itself: Let $t = 0$, then conduct steps *1*$(f)(i)$-(iv), with the exception that $S_{ijk}^{(t+1)} = n_{ij\cdot}^{M} \frac{\pi_{ijk}^{(t)}}{\pi_{ij\cdot}^{(t)}} + n_{ijk}^{E}$ in *Step 1*$(f)(i)$. Denote the resulting estimates after convergence as $\pi_{ijk}^{A(d)}$, where A represents the estimates based on all data.

(i) *Bias in estimates from modified EM algorithm using external data.* For the estimates from the modified EM algorithm that uses the external data, store the vector of biases as $\boldsymbol{\delta}_{A}^{(d)} = \boldsymbol{\pi}^{A(d)} - \boldsymbol{\pi}$.

Step 2: Calculate the average empirical bias of the estimates of the joint distribution based on the external data only as $\frac{1}{D}\sum_{d=1}^{D} \boldsymbol{\delta}_{E}^{(d)}$, and similar for all other types of estimates (complete cases, etc.)

Step 3: Calculate the standard deviation of the estimates of the joint distribution based on the external data only as $\sqrt{\frac{\sum_{d=1}^{D}\left[\boldsymbol{\pi}^{(0)(d)} - \bar{\boldsymbol{\pi}}^{(0)}\right]^2}{D-1}}$, where $\bar{\boldsymbol{\pi}}^{(0)}$ is the vector of means of $\boldsymbol{\pi}^{(0)(d)}$. The standard deviation for the the other estimates (complete cases, etc.) follows similarly.

The results of *Algorithm 2* (two-way interaction model) with data MNAR ($\psi_{\cdot\cdot 1} = 0.55$, $\psi_{\cdot\cdot 2} = 0.3$) are given in Table 3.2. Empirical bias and SD were calculated after $D = 2000$ iterations ($n^E = 300$). Here, $\epsilon_1 = 1.0 \times 10^{-8}$, $\epsilon_2 = 1.0 \times 10^{-7}$ and letting $\boldsymbol{\lambda} = (0, 0.02, 0.2, 0.6, 0.03, 0.07, 0.85)$ determined the population parameters, $\boldsymbol{\pi} = (0.051, 0.063, 0.053, 0.066, 0.094, 0.268, 0.103, 0.302)$.

Here, the initial estimates alone were sufficient to increase efficiency under this model structure, as can be seen in column 6. Specifically, when $n^M = 300$ (i.e., equal to n^E), the estimates from the modified EM given the initial estimates only were

sometimes more efficient than those from the external data alone. However, when this sample size was increased slightly ($n^M = 320$), the modified EM always produced more efficient estimates. As expected, incorporating the external data itself into the modified EM algorithm (column 7) also increased efficiency relative to the external estimates alone. Results from the modified EM algorithm in the aforementioned cases were unbiased, despite the fact that the data were missing not at random. The complete-case analysis (column 5) reflects the bias in the estimates that would be present if the missing data mechanism were ignored.

Table 3.2: Performance of the modified EM algorithm for a two-way interaction (three-way contingency table with no three-way interaction) model with data MNAR. The average empirical bias (SD) for estimates of the distribution of (X, Y, Z) is reported under various estimation methods. SDs in bold represent estimates as or more efficient than those based on the external data alone.

n^M	$\boldsymbol{\pi}$	Population parameters	External data ($n^E = 300$)	Complete cases	Initial estimates from external data	Initial estimates and external data
300	π_{111}	0.051	-1.97e-04 (0.0113)	1.53e-02 (0.0115)	1.42e-05 (0.0117)	-9.16e-05 (**0.0103**)
	π_{121}	0.063	-2.19e-04 (0.0123)	1.83e-02 (0.0135)	-2.69e-04 (0.0123)	-2.43e-04 (**0.0117**)
	π_{211}	0.053	5.97e-05 (0.0115)	1.58e-02 (0.0120)	3.33e-04 (**0.0114**)	1.96e-04 (**0.0103**)
	π_{221}	0.066	1.85e-04 (0.0133)	1.99e-02 (0.0137)	2.38e-05 (**0.0131**)	1.04e-04 (**0.0127**)
	π_{112}	0.094	-4.79e-04 (0.0161)	-8.23e-03 (0.0211)	5.63e-05 (**0.0160**)	-2.11e-04 (**0.0131**)
	π_{122}	0.268	-1.60e-04 (0.0251)	-2.41e-02 (0.0322)	-1.51e-04 (0.0247)	4.65e-06 (**0.0195**)
	π_{212}	0.103	-3.76e-04 (0.0164)	-9.27e-03 (0.0220)	6.06e-05 (0.0168)	-1.57e-04 (**0.0135**)
	π_{222}	0.302	8.67e-04 (0.0261)	-2.77e-02 (0.0339)	-6.88e-05 (0.0265)	3.99e-04 (**0.0207**)
320	π_{111}	0.051	-2.04e-04 (0.0114)	1.56e-02 (0.0115)	-1.16e-05 (**0.0111**)	-1.05e-04 (**0.0101**)
	π_{121}	0.063	-1.51e-04 (0.0126)	1.87e-02 (0.0132)	-2.11e-04 (**0.0126**)	-1.82e-04 (**0.0121**)
	π_{211}	0.053	-1.21e-04 (0.0115)	1.57e-02 (0.0120)	-4.21e-05 (**0.0113**)	-8.01e-05 (**0.0103**)
	π_{221}	0.066	-4.08e-04 (0.0133)	1.96e-02 (0.0132)	- 4.58e-04 (**0.0131**)	-4.34e-04 (**0.0127**)
	π_{112}	0.094	-4.24e-05 (0.0157)	-8.88e-03 (0.0199)	-1.18e-04 (**0.0153**)	-2.66e-04 (**0.0125**)
	π_{122}	0.268	3.50e-04 (0.0251)	-2.45e-02 (0.0315)	1.53e-04 (**0.0241**)	2.49e-04 (**0.0194**)
	π_{212}	0.103	1.99e-04 (0.0165)	-9.16e-03 (0.0213)	2.83e-04 (**0.0165**)	2.43e-03 (**0.0132**)
	π_{222}	0.302	7.58e-04 (0.0253)	-2.69e-02 (0.0321)	4.04e-04 (**0.0250**)	5.75e-04 (**0.0196**)

3.4 APPLICATION IN AN OVARIAN CANCER STUDY

This section considers the application of the modified EM algorithm to data published in Madsen (1976) regarding ovarian cancer. Briefly, a retrospective study was undertaken to assess possible predictors of patients' likelihood of survival past 10 years after treatment. Ultimately, four of the measured predictors were determined to be the most important: stage of tumor at time of surgery (low vs. high), type of operation (extensive vs. not extensive), whether or not the patient received radiation treatment and tumor pathology (localized vs. spread). The outcome was a binary variable that grouped survival as < 10 or ≥ 10 years.

In order to apply the modified EM algorithm in the context of three binary variables, the data have been collapsed over type of operation and pathology. As a result, the measures under study are the stage of the tumor, whether or not the patient had radiation and survival. Of interest is whether, conditional on stage, radiation is associated with survival. Specifically, let X represent radiation; Y, survival and Z, stage. Then, the conditional independence model to be tested is that from Section 3.2.4.1, $X \perp Y \mid Z$. Table 3.3 presents the fully-observed data as provided in Madsen (1976). Analysis of this data using the Mantel-Haenszel test of conditional independence resulted in $p = 0.8957$, thus the test failed to reject the null that radiation is not correlated with survival given stage. The point estimate of the odds ratio was 0.913 (95% CI: 0.494, 1.687).

So as to simulate a situation where data were missing not at random, stage (Z) was made missing at a rate of 0.55 if $Z = 0$ and 0.4 if $Z = 1$, and this data was used as the hypothetical study data from this point forward.

In order to obtain consistent initial estimates for the modified EM algorithm, a random sub-sample of 40% of the data was drawn, and the missing values "recovered"

Table 3.3: Classification of ovarian cancer survival (< 10 vs. ≥ 10 years; $n = 299$) by stage (low/high) and radiation (no/yes).

		Low stage			High stage		
		Radiation			Radiation		
		No	Yes	Total	No	Yes	Total
Survival	< 10	11	20	31	41	77	118
	≥ 10	54	73	127	7	16	23
	Total	65	93	158	48	93	141

(which in practice would have occurred through better attainment of hospital records, e.g.). This data-recovery strategy was discussed in Section 3.1. The 40% sub-sample was considered the external data, while the remaining 60% was the data set subject to missingness.

As shown in Table 3.4, the complete-case (CC) statistics (in red) differ notably from the others, which reflects the bias due to the data being missing not at random. Although this value is still not statistically significant at the $\alpha = 0.05$ level, the statistics obtained from the CC data tell a notably different story to those from all other analyses. The inference changes from a very large p-value to one that is borderline significant, and the point estimate and confidence interval also indicate there could be a significant correlation between radiation and survival, controlling for stage. In the last column of Table 3.4, one will note the width of the 95% CI is smaller for the modified EM algorithm (in bold) than for the external data, indicating an increase in efficiency.

Table 3.4: Results of the Mantel-Haenszel test of conditional independence of radiation and survival given stage for the ovarian cancer study. Point estimates of the OR and associated inference are provided for the true data, external data only, complete cases and modified EM algorithm (which utilized the external data).

	$\hat{\text{OR}}$	p-value	95% CI	CI width
True data	0.913	0.896	(0.494, 1.687)	–
External data	0.931	0.859	(0.340, 2.168)	2.134
Complete cases	2.465	0.067	(1.025, 5.924)	–
Modified EM	0.981	0.993	(0.559, 1.764)	**1.205**

3.5 DISCUSSION

In this chapter, a modified form of the expectation maximization algorithm was presented and applied in the context of contingency table analyses under various model structures. Through algebraic manipulation of the general EM algorithm, the modified EM allows one to estimate most of the terms of the Q-function empirically. The remaining term that must be updated iteratively is not a function of the missing data, and thus there is no concern about the value of the missing data mechanism or the model it is assumed to follow. The assumption of this method is that consistent initial estimates of the model parameters are attainable, which are assumed to come from an external data set. As such, the algorithm is robust to the type of missingness, resulting in estimates that are unbiased and more efficient than the external data alone, even when data are missing not at random and ignoring the missing data would produce biased estimates.

For a set of three discrete variables, (X, Y, Z), where Z is subject to missing values, the computational characteristics of the algorithm were differentiated by whether or not the $n_{ij\cdot}$ are in the set of sufficient statistics for a given model. When they

are, the algorithm is not actually iterative, but rather incorporates information from the data set subject to missingness in a closed-form fashion. In these cases, the initial estimates from the external data alone are enough to provide more efficient estimation when combined with the data set subject to missing values.

When $n_{ij\cdot}$ are not sufficient statistics, the algorithm is iterative, yet the external data (not just the initial estimates) are required to realize an increase in efficiency. This is related to the algorithm's formulation − it incorporates information from the (X, Y) margin into the final estimates. Thus, without the $n_{ij\cdot}$ as sufficient statistics, the method can only increase efficiency through a sheer increase in sample size. In this case, the modified EM is able to combine the external data with that subject to missingness, while still providing consistent estimation.

Because of the algorithm's form, it was recognized that in the case of discrete data, this approach simplifies to a special case of the general EM algorithm: In the data set subject to missingness, all values in the variable with missingness are deleted, then the general EM algorithm is carried out using consistent initial estimates from the external data (and also the external data itself, if available). Of interest, then, is how these two algorithms are related when other types of variables (continuous, e.g.) are considered as opposed to all being discrete. However, this finding may speak to a more general issue regarding the analysis of missing data − namely, without the willingness to make some assumption about the missing data mechanism, an analysis will innately be rudimentary and the gain in information can only be so great.

In light of this finding, this approach may be considered naïve in that it deletes some observed values. However, the trade-off is that it maintains consistent (unbiased) estimates while incorporating additional information, thus increasing efficiency. As a result, if bias is of particular concern in a given study, this approach may be considered advantageous.

BIBLIOGRAPHY

Agresti, A. (2002). *Categorical Data Analysis*. 2nd Ed. Hoboken, New Jersey: John Wiley & Sons, Inc. 3.2.4.2, 3.2.4.2

Anderson, T.W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing, *J. Am. Stat. Assoc.* 52, 200-203. 1.3.4.2

Baker, S.G. and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse, *J. Am. Stat. Assoc.* 83, 63-69. 1.4

Bain, L.J. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics, 2nd ed.* Duxbury. 2.2.3.1

Bishop, Y.M.M. and Fienberg, S.E. (1969). Incomplete two-dimensional contingency tables, *Biometrics* 25, 119-28. 1.4

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press. 1.3.2

Blumenthal, S. (1968). Multinomial sampling with partially categorized data, *J. Am. Stat. Assoc.* 63, 542-51. 1.4

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *J. R. Stat. Soc. B* 22, 302-306. 1.3.3.1

Casella, G. and Berger, R.L. (2002). *Statistical Inference, 2nd ed.* Duxbury. 2.2.3, 2.2.3.1

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1983). Some uses of statistical methods in connection with the nonresponse problem, in *Incomplete Data in Sample Surveys, Vol. III: Symposium on Incomplete Data, Proceedings* (W.G. Madow and I. Olkin, eds.), New York: Academic Press. 1.3.2

Chen, T.T. and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data, *Biometrics* 30, 629-642. 1.4

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: Wiley. 1.3.3.2

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Stat. Soc. B* 39, 1-38. 1.3.4.1, 1.3.4.3, 3.1, 3.1.1

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Stat.* 7, 1-26. 1.3.3.3

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, SIAM, monograph #38, CBMS-NSF. 1.3.3.3

Efron, B. (1988). Three examples of computer-intensive statistical inference, *Sankhyā Ser. A* 50, 338-362. 1.3.3.3, 2.1, 2.2.5.2

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *Am. Statist.* 37, 36-48. 1.3.3.3

Fisher, R.A. (1935). *The Design of Experiments*. Oxford: Oliver & Boyd. 1.3.3.3, 2.1, 2.2.5.2

Fay, R.E. (1986). Causal models for patterns of nonresponse, *J. Am. Stat. Assoc.* 81, 354-365. 1.4

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data, *J. Am. Stat. Assoc.* 77, 270-278. 1.4

Gelman, A.E. and Carlin, J.B. (2002). Poststratification and weighting adjustments, Chapter 19, in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, eds.), New York: Wiley. 1.3.2

Gelman, A.E., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall. 1.3.4

Gimotty, P. and Brown, M. (1987). The effect of imputed values on the distribution of the goodness-of-fit chi-square statistic, *Comput. Stat. Data An.* 5, 201-213. 1.4

Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. 1.3.3.3, 2.1, 2.2.5.2

Hocking, R.R. and Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data, *J. Am. Stat. Assoc.* 66, 65-70. 1.4

Hocking, R.R. and Oxspring, H.H. (1974). The analysis of partially categorized contingency data, *Biometrics* 30, 469-483. 1.4

Holt, D. and Smith, T.M.F. (1979). Post stratification, *J. Roy. Statist. Soc. A* 142, 33-46. 1.3.2

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population, *J. Am. Stat. Assoc.* 47, 663-685. 1.3.2

Ireland, C.T. and Kullback, S. (1968). Contingency tables with given marginals, *Biometrika* 55, 179-188. 1.3.2

Koch, G.G., Imrey, P.B. and Reinfurt, D.W. (1972). Linear model analysis of categorical data with incomplete response vectors, *Biometrics* 28, 663-92. 1.4

Li, K., Meng, X., Raghunathan, T.E. and Rubin, D.B. (1991). Significance levels from repeated p-values with multiply-imputed data, *Statistics Sinica* 1, 65-92. 1.3.3.3, 2.2.5.3, 2.3.4, 2.3.4

Liang, K., and Zeger, S. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* 73, 13-22. 1.3.2, 1.3.4

Lipsitz, S.R. and Fitzmaurice, G.M. (1996). The score test for independence in R $\times$ C contingency tables with missing data, *Biometrics* 52, 751-62. 1.4

Little, R.J.A. (1982). Models for nonresponse in sample surveys, *J. Am. Stat. Assoc.* 77, 237-250. 1.4

Little, R.J.A. (1993). Post-stratification: a modeler's perspective, *J. Am. Stat. Assoc.* 88, 1001-1012. 1.3.2

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd Ed. New Jersey: John Wiley & Sons, Inc. 1.2, 1.3, 1.3.1, 1.3.2, 1.3.3, 1.3.3.1, 1.3.3.2, 1.3.3.3, 1.3.4, 1.3.4.1, 1.3.4.2, 1.3.4.3, 3.1, 3.1.1, 3.1.2

Madsen, M. (1976). Statistical analysis of multiple contingency tables. Two examples, *Scand. J. Stat.* 3, 97-106. 3.4

Miller, R.G. (1974). The jackknife - a review, *Biometrika* 61, 1-15. 1.3.3.3

Oh, H.L. and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse, in *Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography.*(W.G. Madow, I. Olkin, and D.B. Rubin, eds.), New York: Academic Press. 1.3.2

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press. 1.3.4.1, 3.1

Phillips, M.J. (1993). Contingency tables with missing data, *J. R. Stat. Soc. D-STA* 42, 9-18. 1.4

Rao, J.N.K. and Scott, A.J. (1987). On simple adjustments to chi-square tests with sample survey data, *Ann. Stat.* 15, 1-12. 1.4

Robidoux, A., Tang, G., Rastogi, P., Geyer Jr, C.E., Azar, C.A., Atkins, J.N., Fehrenbacher, L., Bear, H.D., Baez-Diaz, L., Sarwar, S., Margolese, R.G., Farrar, W.B., Brufsky, A.M., Shibata, H.R., Bandos, H., Paik, S., Costantino, J.P., Swain, S.M., Mamounas, E.P. and Wolmark, N. (2013). Lapatinib as a component of neoadjuvant therapy for HER2- positive operable breast cancer (NSABP protocol

B-41): an open-label, randomised phase 3 trial, *Lancet Oncol.* 14, 1183-1192. 2.1, 2.4

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *J. Am. Stat. Assoc.* 90, 106-121. 1.3.2, 1.3.4

Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley & Sons. 2.2.3.1

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* 70, 41-55. 1.3.2

Rosenbaum, P.R. and Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling incorporating the propensity score, *Am. Statist.* 39, 33-38. 1.3.2

Rubin, D.B. (1973a). Matching to remove bias in observational studies, *Biometrics* 29, 159-183. 1.3.3.2

Rubin, D.B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics* 29, 185-203. 1.3.3.2

Rubin, D.B. (1976). Inference and missing data (with discussion), *Biometrika* 63, 581-592. 1.2

Rubin, D.B. (1978). Multiple imputation in sample surveys, *Proc. Survey Res. Meth. Sec., Am. Statist. Assoc. 1978*, 20-34. 1.3.3, 1.3.3.3

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. 1.3.3, 1.3.3.3

Stasny, E.A. (1986). Estimating gross flows using panel data with non-response: an example from the Canadian Labour Force Survey, *J. Am. Stat. Assoc.* 81, 42-47. 1.4

Vermunt, J., van Ginkel, J., van der Ark, L. and Sijtsma,K. (2008). Multiple imputation of incomplete categorical data using latent class analysis, *Sociol. Methodol.* 38, 369397. 1.4

Wang, H. (2006). Two-way contingency tables with marginally and conditionally imputed nonrespondents. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison. 1.4, 2.1, 2.2.1, 2.2.3, 2.2.4, 2.2.4.1, 2.2.4.1, 2.2.5.1, 2.3.5, 2.5

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Ann. Stat.* 11, 95-103. 1.3.4.3, 3.1